Summer 5-12-2020

The Potential Link Between Teacher Evaluation and Student Achievement

Darrell Stinchcomb
darrell.stinchcomb@student.shu.edu

The Potential Link Between Teacher Evaluation and Student Achievement

Darrell Stinchcomb

Seton Hall University

Dissertation Committee

Dr. Richard Blissett, mentor

Dr. Christopher Tienken

Dr. Ray Frederick, III

Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Education

College of Education

Seton Hall University

2020

COLLEGE OF EDUCATION AND HUMAN SERVICES
SETON HALL UNIVERSITY

APPROVAL FOR SUCCESSFUL DEFENSE

Darrell Stinchcomb has successfully defended and made the required modifications to the text of the doctoral dissertation for the Ed.D. during this **Summer Semester 2020**.

DISSERTATION COMMITTEE
(please sign and date beside your name)

Mentor:
Richard Blissett — 5/29/2020
Date

Committee Member:
Christopher Tienken — 5-12-2020
Date

Committee Member:
Ray Frederick, III — 5-12-2020
Date

The mentor and any other committee members who wish to review revisions will sign and date this document only when revisions have been completed. Please return this form to the Office of Graduate Studies, where it will be placed in the candidate's file and submit a copy with your final dissertation to be bound as page number two.

Acknowledgements

"He that dwelleth in the secret place of the most High shall abide under the shadow of the Almighty. I will say of the Lord, He is my refuge and my fortress: my God; in him will I trust." Psalm 91

First and foremost I would like to thank God Almighty for all that He has done, is doing, and will do for me. Thank you Lord for hearing my prayers and cries.

I would like to thank Dr. Richard Blissett for picking up the pieces of my work and helping me to shape them into something worthy of reading. I will forever be grateful and thankful for all of your support and guidance. Thank you to Dr. Christopher Tienken and Dr. Ray Fredrick for agreeing to serve on my committee. I respect you and your work and I appreciate your feedback and support.

A special acknowledgment goes to Dr. Elaine Walker, my original advisor who started me on this journey. I sometimes try to imitate your accent when you call out my name. Your breadth of knowledge and calm demeanor helped me to jumpstart this journey. Thank you for your guidance. I pray that you are well and in good spirits.

I could not have done this without the support of my family and friends who encouraged me to not give up even when I wanted to throw everything in the garbage. Thank you to Dr. Lenora Boehlert for being a mentor, coach, and friend. I am truly grateful that you believed in me, hired me, and continue to support me.

It has been a blessing and a pleasure to be a member of Cohort 21 at SHU. I have developed friendships that will last forever. I can truly say that we bonded and continue to support each other in a very special way. I am proud to have worked with such intelligent and caring educators. I pray that God will continue to bless you and your work.

Dedication

I would like to dedicate this achievement to my grandmother, Mary Lee Smith, "Momma," born March 1, 1910. I was born on your birthday and every year we shared a chocolate layer cake, just the two of us. It has been more than eighteen years since you departed this world and I continue to enjoy a slice of chocolate layer cake and shed a tear every time your name is mentioned. I truly appreciate all of the lessons you taught me. Even though I grimaced at the repetition of the stories told, you would be proud to know that I internalized all the intended lessons. You taught me so much and for that I am truly grateful. I owe all that I am and all that I have accomplished to you. I love you, Momma! Rest in Peace!

Abstract

The push for educational accountability and standardization in the United States gained traction with the No Child Left Behind Act of 2001. Uniformity in the curriculum, academic standards, testing, and accountability were some of the requirements that were being touted by politicians, educators, and special interest groups. School districts across the United States were forced to develop systems to prove that teachers were teaching and students were learning. New York State enacted reform legislation under Education Law section 3012-c, which included the Annual Professional Performance Review (APPR) to evaluate teachers and principals. One of the components of this evaluation system consisted of the use of New York State ELA and math scores for students as a means to measure student achievement and was incorporated into the overall ratings for teacher effectiveness.

The purpose of this quantitative study was to examine the potential link between teacher effectiveness in New York State as measured by APPR scores and its possible relationship to student achievement as measured by New York State ELA and math scores. The study sought to examine and establish a definitive relationship between teacher effectiveness and student achievement in New York State as a whole. Some of the essential questions of this research were as follows: What is the relationship between APPR and achievement in ELA and math at the school level when controlling for student characteristics (enrollment, free lunch, reduced lunch, and economically disadvantaged)? What is the relationship between teacher effectiveness and student achievement in ELA and math at the school level when controlling for teacher qualifications (experience and highest degree)? What is the relationship between student achievement in ELA/math and teacher effectiveness (APPR ratings) at the school level?

The study included schools within Orange County, Wyoming County, Westchester

County, Nassau County, and Suffolk County regions in New York State.

The study included a total of 37 school districts, 155 schools, 93,340 students, and 6,915 educators. Data from the 2015–2016 New York State Education Department for both teacher and student scores were used. In 2015, Governor Cuomo issued a moratorium on the use of student achievement scores to calculate teacher APPR scores. Thus, in this study, the teacher APPR scores did not include student achievement scores. This study explored and potentially identified the relationship between teacher effectiveness and students' achievement.

By understanding the relationship between teacher effectiveness and student achievement, individual states, New York, in particular, may be better equipped to direct resources and assistance to school districts that are most in need.

Key words: teacher evaluation, teacher effectiveness, student achievement, accountability, standardization, uniformity, standardized tests, observations, relationship

Table of Contents

List of Tables

Chapter 1

Introduction

The No Child Left Behind Act (NCLB) of 2001 ushered in changes that would forever transform the landscape of public education policy. In an attempt to equalize education across the United States, the laws required uniformity of curriculums, academic standards, testing systems, and accountability—specifically teacher accountability. This firestorm brought on numerous education reform initiatives by state education departments across the United States. Teachers were being held responsible for students' poor performance on international and domestic evaluations that were designed to measure student achievement. Few education issues have received more attention in recent times than the problem of ensuring that elementary and secondary classrooms are staffed with quality teachers (Ingersoll & Collins 2017).

The NCLB reform initiatives compelled school districts across the United States to scramble to come up with systems to prove that teachers were teaching and students were learning. As a system of accountability, New York State eventually enacted education reform legislation that included the Annual Professional Performance Review (APPR) under Education Law section 3012-c to evaluate teachers and building principals. The result was increased testing and assessments in order to provide data to support enforcement of accountability measures for both teachers and principals. A host of initiatives seeking to upgrade teacher quality has been pushed by reformers across the USA and other nations (Ingersoll & Collins, 2017). The world of education was thrust into an era of policies from both the federal and state levels with the expectation of holding educators accountable for what students were learning in the classrooms.

First, there was a high demand for educational accountability. For several decades, there

has been dissatisfaction from policymakers and members of the public regarding teachers' effectiveness and students' achievement. The primary focus of the enactment of the Elementary and Secondary Act (ESEA) of 1965 was to address the educational challenges faced by students who were economically disadvantaged. This changed over time to include an array of issues pertaining to students' performance. The lackluster performance of U.S. students on international evaluations greatly bolstered the credence that students are underperforming (Desilver, 2017). This has mainly been through two arguments: employers' dissatisfaction regarding graduates' unpreparedness in job preparation programs, seeking for, or actually working; and the increasing number of students required to take remedial courses after enrolling in college to catch up (Carnoy & Loeb, 2002).

Second, there is a persistent finding of considerable gaps in student achievement between white and black or Hispanic students, or between economically disadvantaged and advantaged students. These gaps have been documented in various tests including college admissions tests, state assessments, as well as the National Assessment of Educational Progress (NAEP; Tsoi & Bryant, 2015; White et al., 2016). For a number of years, it has been argued that the magnitude of the gaps has remained comparatively constant (Sanders, Wright, & Horn, 1997). The need to reduce these persistent achievement gaps is reflected in the conditions of the (NCLB) Act of 2001 to report student achievement results based on a disaggregated method for various subcategories.

Third, the longstanding belief, according to Hill, Rowan, and Ball (2005), that some teachers do not adequately perform, as far as student achievement is concerned, has also led to demands for teacher effectiveness measures. As a result, various concerns have been raised pressing the public and policymakers to hold teachers and other educators accountable for

students' performance. This has led to a focus on student achievement tests as a cost-effective tool for assessing teacher effectiveness and as a strategy for objectively evaluating the performance of students as an indicator of teacher effectiveness (Kane, Staiger, Grissmer, & Ladd, 2002; Papay, 2012; White et al., 2016).

While the above discusses the importance of accountability, it is imperative to understand the tools used to evaluate teaching and ensure they measure what is intended or are sufficiently linked to student performance. Without a clear-cut connection and effective measures used to examine teaching practices and student outcomes, the issue of accountability or holding anyone accountable is a moot point. Various methods and systems have been employed to determine the degree to which various parties, teachers, students, and schools are committed to the learning process and at the same time determine their individual roles in student achievement. However, only those that are relevant within the context of this study will be briefly addressed.

Problem Statement

Cannell (1987) pointed out two dominant factors that influence student achievement: the assessments employed in measuring the level of performance and the quality of instruction. Initially, measuring or quantifying teachers' effectiveness was a challenge, partly because, until recently, teachers' input as far as the development of curriculum and standards were concerned was minimal. Although student achievement may depend on other factors, teachers' mastery of their roles is a prerequisite.

Despite the growing enthusiasm to develop systems and mechanisms for evaluating how teachers impact the performance of students, often through the use of value-added estimates, systems that integrate student test scores into teacher evaluations have experienced an array of challenges. First, the systems must foster valid and reliable correlations with regard to teachers'

contributions to student learning. Second, the systems must take into consideration the role of teachers who do not regularly teach subjects that are annually tested or do not teach at the grade levels tested (Steele, Hamilton, & Stecher, 2011).

Abrams, Pedulla, and Madaus (2003) opined that it also becomes increasingly difficult to determine teachers' effectiveness on student performance in certain instances, such as those in which the students do not have prior test scores on record or are only enrolled in a class for a portion of the school year. The challenge is how to determine teachers' value-added impact on student achievement when these types of scenarios arise. In some cases, it may be prudent to estimate teachers' value-added impact by using only the achievement of students who are enrolled in classes for a full year or who have prior test scores on record. It would be unfair and problematic to include students without these criteria (Klem & Connell, 2004).

Further, there are specialized institutions that ensure the quality of teaching, frequently through certification, such as the National Board of Certified Teachers (NBCT), National Council for Accreditation of Teacher Education (NCATE), and National Board for Professional Teaching Standards (NBPTS). Ballard and Bates (2008) noted that students with teachers certified through NBCT tend to learn more compared to students in classrooms where teachers do not hold this credential. It, therefore, may be argued that the number of teachers who have been accredited by national certification organizations will undoubtedly raise the levels of student achievement in a majority of schools across the nation. However, in some states, student achievement remains low in spite of teachers being certified by the aforementioned institutions. New York is, indeed, one of these states.

Boyd, Grossman, Lankford, Loeb, and Wyckoff (2009) contend there is a large body of research literature that provides information relevant to understanding how effective teaching

and teacher preparation influence student achievement. However, much of this research is limited in scope and only focuses on the preparation process as opposed to results. In addition, a substantial percentage consists of case study methodologies that fail to describe causal relationships or are not conducive for extrapolation to larger populations (Wayne & Youngs, 2003). This gap in the literature contributes to the need for further quantitative studies, such as the one proposed here.

In addition to the aforementioned, the literature related to New York State supports the need for further research. As an example, Domanico (2018) posited that as far as standardized tests in English language arts (ELA) and math are concerned, most students are not as skilled as the education system in New York State reports. Despite the variation that may occur in many schools as well as between grades, on average over a third of all the students taking ELA assessments in Grades 3 through 8 were deemed *proficient*. While they scored better in math, more than 60% of students still did not perform well.

This variation in students' performance in reading and math year-in and year-out raises concerns with regard to the consistency of the teachers, the teaching practices, and the education system in New York. Domanico (2018) argued that the students have not become any less skilled. Rather, New York's accountability system reflects changes in standards over the years. Ultimately, differences in scoring, as well as the various ways through which tests were administered, have made it difficult to determine student growth in a long-term capacity. The critics of the accountability system in New York State argue that the test scores are not consistent with other measures of student performance, such as the Regents exam or graduation rates.

Fryer (2013) asserted that the introduction of reading and math exams to all Grade 3 through 8 students in New York State occurring in 2006 was a way of complying with federal

policy, but undermined the standardized tests that had been the norm for years. Although New York students were said to have made substantial gains on state tests by 2009, the state had decreased the number of questions students were required to answer in order to pass. In addition, when measured against student performance at the national level, New York State did not demonstrate comparable improvement. In response, the state argued that "cut scores" resulted in the most predominant method for students being deemed proficient. This resulted in a significant drop in student performance. Before the schools could adjust to the new system, the state implemented more changes, in 2013, introducing a new test tied to the Common Core learning standards (Troia & Olinghouse, 2013). The current study explored whether standardized test results can act as a potential link between teacher effectiveness and student achievement.

Purpose of the Study

The purpose of this study was to examine the potential link between teacher effectiveness in New York State and its relationship to student achievement when measured by standardized test scores. The quality of teaching was represented by Annual Professional Performance Review (APPR) ratings, while student achievement was evaluated in terms of student performance on New York State ELA and math tests.

This study is warranted considering daily instructional practices are being revised in order to produce more favorable student outcomes on the New York State ELA and math standardized tests. However, in order to provide effective instructional guidance educational administrators must first understand how the curriculum and daily instruction is being implemented by these changes. This is of little relevance if the link between teacher effectiveness, as indicated by APPR ratings, and student achievement, as demonstrated by

standardized test performance, is not evident. Therefore, in light of the information above, this study is crucial in its efforts to examine and establish a definitive relationship between teacher effectiveness and student achievement in New York State as a whole.

Research Questions

In conducting this study, the researcher sought to answer the following questions:

RQ1: What is the relationship between APPR ratings and achievement in ELA and math at the school level when controlling for student characteristics (enrollment, free and reduced lunch, and economically disadvantaged)?

A. ELA with controls

B. Math with controls

RQ2: What is the relationship between teacher effectiveness and student achievement in ELA and math at the school level when controlling for teacher qualifications (experience and highest degree)?

A. ELA with controls

B. Math with controls

RQ3: What is the relationship between student achievement in ELA and math and teacher effectiveness (APPR ratings) at the school level?

A. ELA without controls

B. Math without controls

Conceptual Approach

According to Ballard and Bates (2008), the way in which schools operate and curriculum is developed throughout the nation has increasingly relied on standardized test results. This has been accompanied by growing pressure from a variety of sources on both teachers and students.

In response, this study examined the potential relationship between content developed by teachers and the quality of teaching and its connection to student achievement. This is relevant when considering that students' performance on standardized achievement tests has been interpreted as a way of reflecting the quality of instruction they receive, as well as the capacity of students to follow instructions.

Some standards, such as the APPR ratings, have been utilized to measure teacher performance and teachers' capacity to impact student achievement. One prevailing theory has been that teachers, in spite of realizing how student achievement can be maximized, have been reluctant to do so in the absence of incentives, rewards, and sanctions (Linn, 2000).

As previously mentioned, it has been established that students instructed by teachers certified by organizations that verify the quality of teaching, such as the NBCT, tend to perform better on standardized tests when compared to their counterparts not assigned to certified teachers (Ballard & Bates, 2008). This implies that teacher effectiveness is a determining factor in standardized test performance for students.

Limitations of the Study

The sampling size was a limitation, only utilizing data from five New York State counties, so generalizability to New York is limited. As only a limited amount of information was available for individual students and teachers, the study relied on how teachers' APPR ratings could predict students' achievement. This limited drawing conclusions for individuals and instead, drawing from the overall scores of teachers in a school and how it related to students' performance in the school itself. Establishing the influence of all external parties, environmental factors, or other possible confounding variables may be a challenge within the context of this study alone.

Delimitations of the Study

The study was confined to the examination of the performance of New York State teachers in relation to the APPR ratings and the influence on student achievement as evidenced by their performance on the New York State ELA and math tests. Since it only included schools within the state of New York, the findings were confined to this particular state. Although the study can be generalized in a number of aspects related to education, the possible variations in education limits its applicability in other states or the generalization of findings across the nation.

Definitions of Terms

Stanford Achievement Test (SAT). A set of standardized tests used to measure the academic achievement of students in kindergarten through Grade 12.

Annual Professional Performance Review (APPR) ratings. A platform aimed at evaluating the efficacy of both teachers and principals based on factors such as performance, student achievement, and student growth. New York principals and teachers are assessed through this platform and at the end of every year rated according to their effectiveness.

Elementary and Secondary Act (ESEA). Act passed under President Lyndon Johnson with the intention of using education as a tool to fight poverty and represented a landmark commitment to equal access to quality education for all. It is presently the largest repository of federal spending on both primary and secondary education.

Minimum Competency Testing (MCT). A standardized exam of rudimentary skills where a passing score indicates that the examined student has acquired the minimum required knowledge and skills in order to either graduate from high school or progress to the next grade.

National Assessment of Educational Progress (NAEP). A platform developed in 1969 with the intention of measuring student achievement across the nation. It is the only national

platform that frequently assesses students' potential in various aspects of learning.

No Child Left Behind Act (NCLB). U.S. Act of Congress enacted in 2001 and signed into law in 2002 that reauthorized the ESEA and included Title I requirements relating to students who are in any way disadvantaged. In 2015 it was replaced with the Every Student Succeeds Act.

Student performance standardized tests. Tests requiring students to answer the same set of questions selected from common criterion and consistently scored, thereby facilitating a comparison of each student with the related performance of others.

Chapter II

Literature Review

The past few decades have shown an emergence of a clarion call to hold teachers accountable. In the spirit of this movement, nations including the United States have reviewed educational policies in a bid to formulate new approaches to standardization and accountability in teaching. With the law now prompting all the states to hold teachers accountable and ensure that the quality of education keeps improving, the need to understand this push has never been more urgent.

As reported by researchers Ciaccio et al. (2017), New York revamped its teacher evaluation system in 2007 by implementing Education Law section 3012-b. It required three factors to be considered when evaluating a teacher: (1) the teacher's use of available student data when providing instruction, (2) peer review, and (3) an assessment of the teacher's performance by the teacher's building principal or other building administrator. Section 3012-b was New York's first step in developing a teacher evaluation system that linked teacher effectiveness to student performance, as it mandated that teacher evaluations be based on analysis of student data and required a statewide evaluation system that linked teacher effectiveness to student performance (Ciaccio et al., 2017).

When looking at the APPR or the current annual performance review standards for evaluating teachers, the assessment is comprised of three components. Forty percent of the evaluation is based on student achievement. This proportion of 40% is then broken into two subcomponents: 20% based on student growth on state assessments and 20% based on other locally selected measures (Ciacco et al., 2017; Moldt, 2016).

While New York garnered national attention for these efforts, which has led to many

changes both in and outside of the classrooms, it continues to go through several revisions in an effort to hold teachers accountable. Other research studies, however, reported that the law actually was not effective at improving accountability or instructional practices, according to educators themselves (Moldt, 2016).

However, any effort to evaluate a method of assessment, without first understanding what prompted its emergence, can only result in a higher likelihood of ineffectiveness or error. In this regard, the current chapter is intended to briefly explore key points in the historical origins of teacher evaluation and accountability, while also examining the connections to accountability to student achievement as indicated in prior studies. This is particularly relevant when considering that a significant focus within the existing body of literature is dedicated to whether or not teacher evaluation ratings accurately and adequately identify quality educators and sufficiently assess effectiveness of faculty (Adnot, Dee, Katz, & Wyckoff, 2017; Alexander, 2016; Johnson, 2017; Kane, Taylor, Tyler, & Wooten, 2010; Medlock, 2017; Taylor & Tyler, 2012).

These efforts at research are often driven by a common motivating factor, which is to enhance student achievement (Adnot et al., 2017; Alexander, 2016; Johnson, 2017). Yet, if teacher effectiveness, as indicated by teacher evaluation ratings, is not empirically linked to student achievement, then any discussion of evaluation accuracy is pointless. The next chapter presents a synopsis of the existing evidence, resulting from studies that sought to answer this question, ultimately identifying a link between teacher evaluation ratings, teacher effectiveness, and student achievement, or the lack thereof, depending on the findings of each individual study.

This chapter discusses both theoretical and empirical sources, while elaborating on some of the most frequently cited studies in the literature, complemented by the inclusion of the most recent studies of relevance. This chapter illuminates potential gaps and limitations within the

current body of literature as it pertains to the possible relationship between teacher effectiveness and student achievement. This leads to later discussion on the significance of the study proposed here, as well as its contribution in relation to the existing literature and its relevance in answering whether differences in schools' overall achievement may be linked to differences in teacher effectiveness, as indicated by APPR ratings.

The sources that comprise this literature review were derived from research publications, peer-reviewed articles, doctoral dissertations, academic journals, and review articles that were accessed through ProQuest and other peer-reviewed or educational databases. In conducting the literature search for the study, the following search terms and key phrases were used: student achievement, teacher accountability, teacher effectiveness, New York State Annual Performance Review, APPR, NCLB Act of 2001, Race to the Top, standardization, standardized tests, teaching quality, teacher evaluation ratings, and student performance. The subsequent results of this search are discussed in the following pages beginning with the brief history of accountability, thereby establishing a foundation and context for this inquiry.

The Emergence of Standardized Testing, Accountability in Teaching, and Teacher Evaluations

In exploring the emergence of teacher evaluations, it is important to understand the evolution of standardized testing and teacher evaluations, as well as their defining features, in order to understand the forthcoming findings of this study, as related to the independent variable, teacher effectiveness as indicated by APPR ratings, as well as the dependent variable of student achievement, represented by ELA and math standardized test results. Scoring teacher effectiveness through APPR ratings or other types of assessments was born out of the need to enhance student achievement and the push toward standardized testing (Beyer & Johnson, 2014;

Medlock, 2017). In light of the aforementioned, this historical synopsis first begins by defining the concept of standardized testing.

Definition of Standardization in Testing and Accountability in Teaching

Cramer, Little, and McHatton (2018) defined a *standard* as a value or a metric. Framed differently, a standard is an instrument used as an indicator of another. Thus, standardization is a process of determining what metric or value can serve as an indicator of another. In the context of education, standardization would, therefore, refer to the political process of making various units use the same measurements and or outcomes (Cramer, Little, & McHatton, 2018).

Traub and the Canadian Education Association (1994) asserted that the concept of standardization in and of itself implies that all participants' resulting scores may be compared one against the others, because standardization is about uniformity of measurement, not the measurement itself. They continued to explain that a standardized achievement test is typically designed for a predetermined context and involves a method of implementation that ensures it is consistently administered to all student groups in the same manner (Traub & Canadian Education Association, 1994). Scoring is also executed in the same way, regardless of the setting, who administers the test or who oversees it, thereby producing scores that are conducive to comparison in an individual capacity or in an institutional capacity (Traub & Canadian Education Association, 1994).

Good (2008) ascertained that standardized tests are administered in the same consistent manner for all examinees. The content is also the same for all individuals, irrespective of their race, age, gender, sex, or any other functional and personal attributes. Hence, the testing environment and content of standardized tests remain constant at all times (Ballard & Bates, 2008; Good, 2008).

Mathison and Ross (2013) defined accountability as a concept related to authority. According to these researchers, accountability refers to those who possess authority and how it is exercised (Mathison & Ross, 2013). Snowman and McCown (2014) asserted that accountability exists when one is asked to explain and justify his or her actions to one or more parties who have a stake in the task. The researchers drew a parallel to students and teachers, illustrating how the two related in terms of authority. However, the question still remained as to how these concepts emerged within the context of education. Therefore, a brief historical overview is presented in the next section.

History of the Standardization Movement in Teaching

In discussing the concept of standardized testing as it applies within the American field of education, Hamilton and Koretz (2002) pointed out that the current test-based accountability efforts in the United States were in no way novel or innovative. According to these scholars, what was seen as a national push for standardization and accountability could be observed in policies formulated over a century ago. In an effort to prove this assertion, the two presented a brief history of large-scale assessments dating back as early as the 1800s. From the middle of the 19th century forward, schools utilized these tests to compare teachers, as well as to determine curriculum efficacy.

In 1923, stakeholders developed the Stanford Achievement Test (SAT), which was designed for elementary school students and inspired the use of formal and group-administered batteries in assessing a range of academic skills across the field of education (Freedheim, 2003; Hamilton & Koretz, 2002). Although Freedheim (2003) points out that American schools began using achievement testing in the early 1920s, the author also acknowledged that there were tests for specific competencies already in use before the 1920s, such as spelling tests. Two years

later, the Iowa Test of Basic Skills (ITBS) was developed (Freedheim, 2003). Unlike the SAT, ITBS was designed with older students in mind.

Hamilton and Koretz (2002) explained that the 1960s witnessed a significant evolution of large-scale testing programs. During this period, Congress developed the NAEP, thereby requiring an assessment of students' achievement in various subjects, particularly emphasizing civics, geography, science, mathematics, history, reading, and writing (Beatty, Educational Testing Service, & National Center for Education Statistics, 1996). Also during this decade the federal government established ESEA in 1965 (Beyer & Johnson, 2014).

At the time, ESEA served as a way in which the administration could exert its influence on education (Beyer & Johnson, 2014). Since its formulation and implementation, Beyer and Johnson (2014) observed that ESEA has gone through several revisions, specifically, five stages in its journey. After being enacted in 1965, it was revised in 1978, and in 1981, under the title of Education Consolidation and Improvement Act. In 1988, Congress further reviewed the act, resulting in the birth of the Hawkins-Stafford Elementary and Secondary Education Improvements Act. Finally, after another review in 1994, NCLB was enacted in 2002. Of all the revisions, NCLB has become the most contentious, predominantly based on three main factors. The first involves its emphasis on accountability measures and student achievement as captured in Title I of the original Act. The second involves its emphasis on the need to have highly qualified teachers, while the third involves issues related to charter schools, parental choice, and innovative programs (Beyer & Johnson, 2014). NCLB was also the federal government's move into accountability. Many states were already equipped with various forms of test-based accountability.

Many of the goals inherent in the acts have led to a greater reliance on standardized

testing as a means of evaluating teacher effectiveness and student achievement (Beyer & Johnson, 2014). One way of achieving this is through the use of exit exams, as described by Fuller and Henne (2008). These scholars asserted that the history of exit exams dated back more than three decades, with a significant number of states adopting Minimum Competency Testing (MCTs) at the end of the 1970s and at the dawn of the 1980s. Statistically, the number of states using MCTs increased from two percent in 1973 to 34% in 1983. Although MCTs were intended to ensure that high school graduates had mastered basic skills, ultimately these tests served as a transition from large-scale assessments to using assessments aimed at holding schools accountable (Fuller & Henne, 2008).

Mertler (2007) explained that MCTs created a new purpose for these tests, evaluating the performance of both teachers and students, thereby leading to a measure of teacher accountability. In this regard, tests began to be designed with respect to this frame of reference, serving as a tool for improving educational practice. According to Mertler (2007), the emergence of MCTs served as the inspiration behind data-driven instruction.

Teacher Evaluation and Accountability

While the aforementioned sources shed light on how standardized testing for students led to the emergence of teacher accountability, this view has been confirmed by other sources in the literature. More specifically, Ruiz-de-Velasco (2005) asserted that the education reforms of the 1980s and 1990s have continued to influence policy, even in the 21st century. One of the ways in which this influence is evident pertains to the contemporary emphasis on holding teachers accountable for student performance (Ruiz-de-Velasco, 2005).

However, Ruiz-de-Velasco (2005) does acknowledge that the reforms of prior decades may differ from those in the present era. Yet at the core, Ruiz-de-Velasco (2005) observed that

20th century education reforms were calling for standardization aimed at ensuring that all students had equal access, including English as a Second Language students. Earlier legislation, such as the Emergency Immigrant Education Act (1984), the Bilingual Education Act (1968), and ESEA (1964), indicated the origins of an education standardization movement between the 1960s and 1980s. Today, however, the aim of these reforms has shifted focus to an emphasis on performance outcomes. The primary motivation behind this emphasis and the overall push for standardization is driven by an effort to continually increase the quality of education through the use of student testing for promoting teacher effectiveness (Velasco, 2005).

Ruiz-de-Velasco (2005) offered additional support for the origins of the current movement toward teacher effectiveness, occurring in the mid-1980s. He referenced the work of the National Commission on Excellence in Education, which published a 1983 report titled *A Nation at Risk*. Specifically, this report called for new student tests, more effective instructional frameworks, and higher curriculum standards. This translated into a focus on holding teachers accountable using new student assessments (Ruiz-de-Velasco, 2005).

This emphasis on teacher effectiveness in education continued to gain momentum in the 1980s and 1990s, according to Seifert and Vornberg (2002). This focus was particularly evident in the political sphere, as indicated by the G.W. Bush administration's commission of a study targeted at evaluating the progress of students and their level of achievement. Although the study ultimately revealed positive factors as well as areas in need of improvement within the nation's education system, the G.W. Bush administration did not use the findings to create any new educational policies. The underlying interest in the report was driven by allegations from lower socioeconomic communities of these local schools failing to sufficiently educate their students (Seifert & Vornberg, 2002). This possible educational failure in poverty-stricken

communities was an ongoing area of inquiry, leading to further exploration of the education system in the Clinton administration.

Salvia, Ysseldyke, and Bolt (2010) shed light on how the Clinton administration continued the focus on holding teachers accountable for educational outcomes in lower socioeconomic areas. One way in which this was evident was through the mandated adoption of comprehensive accountability systems under Title 1 of ESEA. In light of allegations that schools with large student populations from ethnic minorities or lower socioeconomic families set educational expectations that were below average for their students, President Clinton required all states and the schools within them to meet minimum performance standards for all students, regardless of background. Schools had no choice except to develop performance-based accountability systems. Moreover, some schools began using test scores as a means of assessing principal or teacher effectiveness (Salvia, Ysseldyke, & Bolt, 2010).

Bjork (2015) observed that efforts to make schools and teachers, in particular, more accountable began to significantly increase in January of 2002. During this year President George Bush signed NCLB into law. At the core, NCLB aims at ensuring that every child has an equal, fair, and considerable opportunity to attain a high-quality education. Citing President Bush, Bjork (2015) explained that the administration aimed to end "the soft bigotry of low expectations" (p. 20). In alignment with this goal, the Bush administration promised to: (1) see to it that all students demonstrate improved achievement; (2) every student meets challenging state academic standards; and (3) teaching effectiveness would be strengthened and improved. More specifically, the administration aimed at ensuring that every student would have reached proficiency standards in mathematics and reading by 2014 (Bjork, 2015).

Another source relevant to this discussion is that of Sunderman, Kim, and Orfield (2005),

which specifically elaborates on how NCLB served to enhance standardization and accountability in schools. The researchers reported that within the realm of education state autonomy led to vast differences. NCLB was intended to put an end to this by limiting state autonomy in the context of education (Sunderman, Kim, & Orfield, 2005).

Specifically, NCLB served to achieve standardization and hold teachers accountable through three basic mechanisms. The first aimed to create consistency in education and restrict the variations in state educational quality by expanding the role of federal government within the realm of education. This entailed the Department of Education identifying failing schools, conducting research to discover underlying causative factors, and introducing potential remedies. NCLB also dictated a timeline for proposed changes and mandated state participation in the NAEP (Sunderman et al., 2005). The primary function of NAEP is to serve as an index of student performance.

Second, NCLB allowed for the establishment of district and state systems that compared school performance on the basis of student achievement. The act focused on improving schools as opposed to improving the achievement of individual students. Hence, NCLB shifted focus from whether the implementation of programs was successful to whether student achievement maintained a positive trajectory (Sunderman et al., 2005).

Third, according to Sunderman et al. (2005), NCLB reassigned local authority and, in lieu of local departments of education, delineated state education agencies. Funds from the federal government would be handled through these agencies rather than more local boards. This restructuring shielded education at the state level from being under the authority and capture of local politicians. Also, state education agencies could determine what constituted proficiency, even if school boards did not agree (Sunderman et al., 2005).

As a result of the historical efforts at standardization and the standardized testing created for assessing student achievement, teacher effectiveness emerged. Standardized tests not only evaluated student achievement, but also served to assess teacher effectiveness in the contemporary field of education. And, as such, they serve as a tool for promoting teacher effectiveness. However, student performance as an adequate indication of teacher efficacy is an issue that remains to be seen and is at the core of this study. It is equally important to have a sufficient understanding of evaluations and observations, as discussed in the next section of this review.

Teacher Evaluations

The use of teacher evaluations has been part of the field of education as early as the 1700s. After the Industrial Revolution there was a need for a more experienced and educated work force. Schools were formed so that children and adults could get better jobs. The formal instruction of students in schools established the need to supervise the instructional practices taking place. Supervising and observing teachers was initially the responsibility of the clergy and business leaders. Clergy was the preference, however, because of their teachings in the church and their education background. Marzano & Livingston (2011) posited that the teacher was considered a servant of the community. Teachers carried the ideals of a democratic education, and a democratic education was necessary for the creation of an educated and well-informed populace (Schneller, 2017). With no formal agreement as to the importance of pedagogical expertise, the quality and type of feedback to teachers was highly varied (Marzano & Livingston, 2011). School systems continued to develop, and the need for educators with pedagogical expertise continued to grow. A shift in education began to take place during the 1800s and 1900s with the development of normal schools in the New England states—

specifically, Massachusetts and Connecticut. This shift also prompted a particular interest in the preparation of teachers. These new normal schools were viewed as early teachers' colleges and established the need for educators to be experts. Two ideas for education also developed in the 1800s. John Dewey saw democracy as the conceptual underpinning of human progress (Marzano & Livingston, 2011). Students were viewed as interactive learners and functioning citizens of society. Teacher observations focused on a student-centered environment and the teacher taking on the role of facilitator. Around the same time, the work of Frederick Taylor on scientific management began to influence the work of educators like Edward Thorndike and Ellwood Cubberley, where measurement of behavior played a significant role in schools. While Cubberly's approach centered on the use of data to make decisions, Dewey's focus remained on educational goals and citizenship (Marzano & Livingston, 2011).

The Soviet Union's launch of Sputnik in 1957 put America under scrutiny with regard to science, math, and technology. The Soviet Union's advancement in science and technology made the need for teacher accountability greater. In response to the launch of Sputnik, schools adjusted their curriculums to offer higher levels of science and math classes. The increased demands for teacher accountability shifted the educational focus to the development of teacher skills and the supervisor's role in learning. The clinical supervision model introduced in the 1970s required the teacher and supervisor to plan, observe, analyze, and discuss the teacher's practice (Robinson, S.B., 2020). Teacher evaluation systems such as the Marzano Focused Teacher Model and Charlotte Danielson's Framework were used to observe teachers and continued to emphasize classroom organization and management practices. These models provided a measurement system so that teachers' performance in the classroom could be quantified.

While the results on state standardized tests measured student achievement, teacher evaluation scores were mainly being derived from the supervisor's observations, denoting processes, and still no focus on student learning outcomes. Grissom, J. A., & Youngs, P. (2016) asserted that classroom observations have strong face validity because they assess "process," or teaching variables, not student outcomes, which may feel distal from teachers' work. Federal, state, and local authorities continued to advocate for a multi-tiered structure for evaluating teachers, which would include student achievement scores and observations. Education policy makers began to experiment with the use of student data outcomes from standardized tests to measure teacher performance. Value Added Modeling was developed and education policy makers promoted it as a useful tool for evaluations.

Measuring Growth through Teacher Evaluation Systems

Student Growth Models (SGMs) and Value Added Modeling (VAMs) are but two ways student data outcomes inform teacher effectiveness. SGMs used in some districts utilize a methodology that describes student achievement by examining individual student growth as compared with similar student profiles. SGMs indicate academic growth and can predict student performance. A variable of academic growth and student performance is teacher effectiveness. Monitoring teacher effectiveness is the function of teacher evaluations. Teacher evaluations can influence teacher effectiveness, which can, in turn, influence student growth. It is this phenomenon that intersects Student Growth Models with Value Added Models. Although the question is not fully settled, these models find that teachers vary substantially in their contribution to achievement growth and that exposure to high value-added teachers has measurable positive effects on students' educational attainment, employment, and other long-term outcomes (Bitler et al., 2014).

The Value Added Modeling (VAMs) was introduced as an effort to measure the teacher's contribution to student learning over time by comparing student performance results of test scores from a pretest and a post test. The use of VAMs attempted to assess the broader construct of teacher quality by measuring a student-specific construct, growth in learning, or test performance (Grissom, J. A., & Youngs, P., 2016). Educators, politicians, and interest groups seeking to further develop teacher accountability began to advocate for the use of VAMs as a measurement tool. Federal, state, and local authorities began a push toward a multi-tiered structure for evaluating teachers, which would include student achievement scores.

Many state education departments required school districts to incorporate some form of VAMs into their teacher evaluation systems, but it did not come without controversy. Many education policymakers noticed the weakness in using VAMs, such as the reliance on state assessments that might not accurately capture the type of learning that was considered to be important (Grissom, J. A., & Youngs, P., 2016). New York State was no exception to the protest against the use of VAMs and standardized test results to evaluate teachers.

Facing a revolt from parents and teachers, Governor Cuomo and the New York State Board of Regents issued a moratorium on the use of standardized test scores in the teacher evaluation systems in 2015. The field of education continues to garner criticisms for the operation of schools, compensation, standardized tests, and teacher evaluations, to name a few.

Teacher Evaluations and Student Achievement

As previously mentioned, within the context of this study student achievement is represented by standardized testing outcomes. Whether or not this variable has a relationship with teacher evaluations, in this case APPR ratings, teacher effectiveness is at the core of this research. Earlier studies have explored these potential relationships in different capacities

through the use of varying methodologies. In spite of these efforts, the inconsistency among research findings and the conflicting results have led to the need for additional studies, such as this one, that continue to examine this question in the hope of finding more definitive answers. As such, a brief synopsis of the current inventory of relevant literature is presented next.

Evidence of a Relationship Between Teacher Evaluations and Student Achievement

In light of the ongoing emphasis on teacher effectiveness, many published studies and dissertations have set out to examine various factors related to this issue. Among these, the dissertation published by Johnson (2017) focused on the potential relationship between the effectiveness of teachers and student growth. The study was facilitated in response to the TEACHNJ Act, which mandated that teacher tenure would, at least in part, be determined by the teachers' evaluation ratings, in an effort to improve the level of teaching and, in turn, enhance student growth as a result. The quantitative analysis involved several variables, including many at the school level, the teacher level, and characteristics of the students. One of the predominant questions at the core of this study was identifying how student growth might or might not be influenced by a teacher's effectiveness, as represented by their practice score or evaluation rating (Johnson, 2017).

Johnson's (2017) sample of participants were all relative to New Jersey; the teachers participating were employed to teach Grades 4 through 7 in either language arts (N = 149) or mathematics (N = 145) from thirty participating schools. Ordinal regression was then utilized as the analytic method for examining the possible relationship between teacher characteristics and student growth, ultimately determining that a positively correlated relationship existed. The researcher found that as teacher ratings increased, so did student growth, regardless of the urban setting and ethnic composition of the student sample (Johnson, 2017).

These conclusions reaffirm the earlier findings of the Bill and Melinda Gates Foundation (2013) in which the Measures of Effective Teaching (MET) entailed a composite score for evaluating teachers. This weighted measure created an accurate assessment of teaching efficacy, devoid of the bias associated with an overemphasis on any one factor. Within the MET study, composite scores for teachers were tested for a relationship with student achievement as indicated by state standardized tests. Using correlation and regression analyses, it was found that a teacher's composite score could accurately predict the level of student performance associated with them the following year (Bill & Melinda Gates Foundation, 2013).

In addition, when examining individual student performance from one academic year to the next, students were randomly assigned to a teacher categorized as *effective* or *less effective*. Those assigned to the effective teacher group ultimately performed better than expected, according to their prior test performance, while those assigned to the less effective group performed worse than expected (Bill & Melinda Gates Foundation, 2013). These findings provide credibility to earlier studies such as that of Papay (2012), in which a correlation or association was statistically identified between student achievement and teacher evaluation ratings.

In another study, published by Taylor and Tyler (2012), teacher evaluations improved student performance, but as a function of the evaluation process itself. In other words, it was found that after teachers underwent the evaluative process, their students scored higher on standardized tests the following year. Specifically, students received scores that were .11 standard deviations higher than the teacher's students in the year before the evaluation took place. As a result, this study indicates another way in which student achievement may, in fact,

be linked to teacher evaluation ratings in that teachers invested a greater effort after undergoing the assessment.

Perhaps one of the most compelling studies was that of Chetty et al. (2010) in which teachers' impact on student achievement was assessed with regards to the gains made by students in standardized test scores. In doing so, the researchers explored the effect that occurred, if any, after a teacher or teachers with highly effective or strong track records left one school and worked at another. Subsequently, it was found that when those teachers categorized as having more effective track records with students actually left a particular school, the performance of students in that grade level worsened overall. Conversely, when a highly effective teacher joined the faculty at a new school, the performance level of students in that new school were elevated (Chetty et al., 2010).

Although the findings of this study were not definitive, they do provide a persuasive illustration for the impact more effective teachers have on student achievement and, reciprocally, how standardized test scores may, indeed, be a good indicator of teacher efficacy. In fact, these researchers further elaborated that while grade-level performance of students changed in response to a teacher leaving or joining a school, the performance of students in other grades remained unchanged, thereby enhancing the credibility of the findings realized within this research endeavor (Chetty et al., 2010).

Studies Producing Alternate Findings

While the aforementioned studies serve as proof of the relationship between teacher evaluation ratings and student achievement, the literature was also rife with studies that produced conflicting results. Among these, the dissertation published by Alexander (2016) focused on teachers and students within the state of Illinois. This study also examined standardized test

outcomes, regarding math and reading, specifically, which is similar to the study presented in this paper. However, distinct from this study, Alexander (2016) utilized the Measures of Academic Progress as the instrument, which measured student outcomes.

Alexander's (2016) final participant sample was derived from seven elementary schools, but featured only fifth-grade students (N = 317) and teachers employed at the same grade level (N = 19) for the 2015–2016 academic year. A correlation analysis was then implemented for testing the potential relationship between teacher evaluation ratings and student math and reading test performance, respectively. As a result, the researcher reported no statistically significant relationships between any of the variables tested (Alexander, 2016).

Perhaps even more interesting was that examining the correlation outcomes more closely, negative correlations were realized in each case with a Pearson's *r* of -.074 (p = .188) and -.103 (p = .069) for math and reading, respectively. Therefore, although the subsequent relationships were not significant, as teacher effectiveness improved, as measured by evaluation ratings, student performance actually worsened. These outcomes persisted, even in spite of the fact that the study attempted to control for potential confounding variables by excluding students with excessive absences or those who were included in special education, as indicated by an individualized educational plan (Alexander, 2016).

In another research endeavor, Medlock (2017) focused on a high-performing state regarding student standardized testing outcomes in order to examine a potential underlying causation for the ethnic variation that persisted. More specifically, an achievement gap existed between Caucasian students and their African-American counterparts within the state of North Carolina. In this instance, the standardized test used as the instrument of measurement was the state end-of-grade test on mathematics for 8th grade students for the 2014–2015 and 2015–2016

academic years (Medlock, 2017).

Ultimately, the mixed methods analysis revealed that teacher evaluation ratings were not a predominant indicator of student achievement nor were student characteristics responsible for the distinct gap between students of different ethnic backgrounds. Instead, after quantitative methods combined with qualitative interviews were analyzed, it was found that teachers' lack of interest in understanding cultural factors may prove influential, as well as differing learning styles, that were the primary drivers behind the gap that continued to plague an otherwise high-performing district (Medlock, 2017).

In a similar capacity, Berliner (2013, 2014) found that teacher evaluations did not predict student performance, but socioeconomic class was an influential variable. More specifically, students of a higher social class were associated with increased numbers of students who passed while lower socioeconomic students were associated with higher fail rates (Berliner, 2013).

Finally, the study of Forman and Markson (2015) examined the potential relationship between teacher evaluation ratings, represented by APPR ratings as in the current study, and student achievement within the state of New York. Other factors taken into consideration included per pupil spending, attendance rates, and poverty. Student achievement was represented by Grades 3 through 8 ELA and mathematics assessments, derived from Nassau and Suffolk counties, totaling approximately 60,000 students and data from 30,000 teachers (Forman & Markson, 2015).

Somewhat similar to the findings of Berliner (2013), poverty was negatively correlated with student achievement, as indicated by standardized testing outcomes, thereby indicating that as poverty increased, student scores decreased (Forman & Markson, 2015). In fact, this was such an influential factor that on both the ELA and math assessments this variable accounted for

over 60% of the variation in student scores. In contrast, APPR ratings for teachers rated as *highly effective* were positively correlated with student achievement, indicating that as the number of *highly effective* teachers went up, student test scores went up as well. The greater influence was found to be realized among ELA scores and, even in this instance, teacher effectiveness was only found to be responsible for 12.53% of the variation in student scores (Forman & Markson, 2015).

Another interesting and conflicting finding emerged in that the percentage of teachers rated *effective* had a unique effect on student performance, presenting as negatively correlated with student standardized test outcomes. In other words, as the percentage of teachers rated *effective* increased, the performance of students actually went down. These results were statistically significant for both *highly effective* and *effective* teachers. In essence, the authors note that there may have been underreporting of *ineffective* teachers and, therefore, many teachers who were rated as *effective* had a negative impact on student performance, because they actually were *ineffective* (Forman & Markson, 2015).

In response to the conflicting findings within the literature, many researchers have attempted to identify possible underlying reasons or discover if there are problems inherent in the use of teacher evaluation ratings as they relate to teacher effectiveness and student achievement as a whole. These findings are not only relevant in that this study may or may not discredit these possible concerns, but also in that they may present as possible limitations of the current study, dependent on the outcomes that are realized. The relevant literature related to these concerns is presented next.

Potential Issues with Teacher Evaluation Ratings

While the studies previously presented attempt to answer whether or not student

achievement is linked to teacher evaluation ratings, the studies in this section attempt to address *why* there may be an issue with teacher evaluation ratings in this context. Marshall (2013) reports several factors that were in conflict with the use of teacher evaluation ratings to predict student achievement or as an indicator. First and foremost, this researcher asserted that the student tests were simply not designed with the purpose of assessing teachers. In this type of value-added assessment, a teacher's data would need to be collected for a period of at least three years in order to achieve any accurate results. Failing to do so would produce findings that were biased because of confounding factors or extraneous "noise" (Marshall, 2013).

Darling-Hammond, Amrein-Beardsley, Haertel, and Rothstein (2012) is a frequently cited study within the literature in which teacher evaluations are discussed. One of the critiques of using value-added assessments in this context was that a significant percentage (25-45%) of teachers rated *ineffective* or *less effective* in one year were frequently rated *highly effective* the next year. Similarly, the converse was true in that *highly effective* teachers in one academic year were often rated as *less effective* in the subsequent year. As a result, the variability of teacher ratings appears to lack consistency and, therefore, provides information that may be meaningless from one year to the next (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012).

In addition, Darling-Hammond et al. (2012) also purported that wide variations in a teacher's performance might occur simply as a function of the students he or she was assigned during any particular semester or academic year. And finally, the assertions of Darling-Hammond et al. (2012) confirmed those of Marshall (2013) in that teacher evaluations failed to account for or control for the many extraneous factors that might also impact student performance.

In fact, Darling-Hammond et al. (2012) offered a substantial inventory of other factors

that either contribute to or impede gains in student achievement, dependent on the individual, including school level factors such as class size, resources, curriculum, and availability of tutoring. The student's family and household environment may present a challenge or pose as a benefit in terms of support as well as the peer group or school culture. Compounding these influences, an individual's specific needs, preferred learning style, strengths and weaknesses, psychological and physical health, as well as attendance, inevitably made an impact. Finally, a student's prior learning experience will likely prove influential, as the influence of teaching in former grades is cumulative and will undeniably have an impact on the student's current performance (Darling-Hammond et al., 2012). In light of these varied influences, standardized testing outcomes may not be an accurate assessment of a teacher's impact on student performance, without controlling for these additional influencing variables (Darling-Hammond, 2013, 2014; Darling-Hammond et al., 2012).

Later publications by Darling-Hammond (2014) state that the specificity of standardized testing for students is intended to measure grade level skills. As mandated by NCLB, these tests do not assess higher skills, nor do they evaluate prior learning skills, thereby falling short of actually measuring the achievement level of a student and, instead, simply testing whether or not they have mastered a set of basic, current skill sets. In the end, the use of teacher evaluation ratings that involve student data from standardized tests may lower, not improve, the quality of teaching, as educators may focus on specific content that will be presented on the test in an effort to improve student performance. The weakness inherent in this approach is that "teaching to the test" often means neglecting other necessary skills or topics simply because they are not included in the standardized test content (Darling-Hammond, 2014).

In general, teaching to the test is a frequently mentioned criticism of linking teacher

evaluation ratings to student achievement. In an earlier study, published by Boyd et al. (2011), experienced educators had students who performed better on their standardized tests than students who were assigned to less experienced teachers. However, the researchers warn that this was not indicative of higher quality teaching or the students having learned more in the experienced teacher's classroom. Instead, they assert that it is simply an indication that experienced teachers are better equipped to gear their curriculum toward content that will be represented on the standardized tests, thereby teaching to the test, so to speak (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2011). Similar concerns were later published by Green, Baker, and Oluwole (2012), Baker, Oluwole, and Green (2013) and more recently by Ciacco et al. (2017).

Franco and Seidel (2014) reiterated the concerns of Darling-Hammond et al. (2012) in that many factors influence student achievement, extending beyond the effectiveness of a teacher or strength of a school's faculty. Looking at an urban school setting, these researchers sought to uncover those variables that may or may not influence student achievement when the ethnic composition, possible socioeconomic status, and other demographic characteristics are not typical compared to the many suburban schools featured in the inventory of literature. Once again, when using value-added measures for assessing teacher effectiveness combined with student achievement as a measure of teaching efficacy, there were influential factors at the student, teacher, and school level (Franco & Seidel, 2014).

Franco and Seidel (2014) indicated that these confounding variables make it difficult to discern how much student growth may be a reflection of teacher effectiveness in and of itself. Many of these factors were also cited in earlier studies, including socioeconomic factors, the student's progress in the prior academic year, as well as the level of parents' education. A new, but seemingly obvious factor that is worthy of mention is a student's motivational level (Franco

& Seidel, 2014).

An article published by Ciacco et al. (2017) specifically examined this issue as it pertains to the state of New York and APPR ratings, in particular. These researchers reflected many of the earlier concerns of teacher ratings within the literature, applying them to APPR rating scores, including the evaluation's lack of reliability in that too many additionally influencing factors may contribute to student achievement outcomes. This is once again compounded by the annual nature of the evaluation in that the potential for bias or confounding factors associated with short-term use may be mitigated when the data is analyzed in a long-term capacity or as aggregate data (Ciacco et al., 2017).

Ciacco et al. (2017) also cited the negative, yet unexpected and unintended consequences that often result when APPR or similar evaluation ratings are used. Among these, the authors explained that financial outcomes may emerge that negatively impact teachers, students, and the school as a whole (Ciacco et al., 2017). In fact, a 2010 study published by Baker et al., suggested that factors beyond a teacher's control may impede student achievement in the lower socioeconomic areas, including characteristics of the students. Exceptional teachers may be deterred from working in the neediest schools because of the negative impact student performance will have on their evaluations, particularly in light of the reality that it may have little to do in reflecting the actual quality of their teaching (Baker et al., 2010).

Chapter Summary

In summarizing the review of the literature presented, it is clear that accountability and standardization are not new phenomena faced by the American education system. Through the decades, the pressure to have teachers held accountable using student achievement has undergone several transformations. Initially, people believed that the introduction of

standardized testing for students would result in inequality in schools, especially for students of an ethnic minority. Whether or not this has occurred is beyond the scope of this study, it is unfortunately evident that the achievement gap between Caucasian students and minority students still persists. In some cases, scholars have observed that the 'White-Black achievement gap' has widened over time, even following the implementation of NCLB, Hanushek & Raymond (2004).

Nevertheless, advocates of teacher effectiveness ascertained that creating competition in schools would elevate both teacher and student performance, motivating teachers and students to invest a greater effort toward achievement. At the same time, the federal government introduced the concept of rewarding top-performing teachers and punishing low-performing ones. Because of this policy, as well as other factors associated with accountability, many unintended consequences of teacher effectiveness and its use in connection with student achievement have emerged and warrant attention.

Among these is the reality that if teacher effectiveness ratings (APPR) negatively affects the motivation and morale of teachers, there is no question that the quality of teaching and learning outcomes will be affected. At the same time, if the pressures associated with APPR measures discourage people from joining the teaching profession and incentivize others to leave, this poses a threat to the education field. Additionally, the use of standardized test scores and their relationship to teacher assessments may, in fact, dissuade highly effective educators from accepting employment at schools with more challenged students or greater populations of poverty, as these are often cited as influencing factors.

These unintended outcomes are relevant in that the impact on teachers must be considered and weighed against the benefits of linking teacher effectiveness and student

achievement. The failure to do so can result in outcomes that undermine the very reasons for implementing such policy in the first place. The use of such measures may actually detract from the quality of teaching and impede student achievement rather than improve it.

In further support for additional studies, the research on the relationship between teacher effectiveness and student achievement has mixed results at best. There are a host of empirical studies that have proven accountability policies, including NCLB, have had notable positive impacts on standardized testing outcomes and NAEP scores. Others have maintained that NCLB has generated negative impacts, not only on student achievement, but also on education as a whole.

These findings are further compounded by the inventory of studies that conclude the relationship between teacher effectiveness and student achievement is dependent on several variables. These may include, but are not limited to, state and school constitution related to the ethnic composition of the student population, as well as socioeconomic status and a myriad of other factors at the student and school level.

In addition, many research endeavors have explored the relationships of accountability policies in several states, rather than the connections in a particular state or local region. Of those that focused on a particular state, New York is not typically the setting for the study, thereby failing to examine the relationship of these variables within the context of the unique urban and ethnic composition of the student population at the core of the study proposed. Finally, many studies within the literature are not conducive to extrapolation as the use of correlation may indicate a relationship, but not causation, or the analyses involved failed to control for conflicting or confounding factors.

In light of the aforementioned, the current study recognizes these inherent weaknesses

within the existing inventory of literature and addresses these shortcomings. It attempts to fill the gap within the literature by examining the potential relationship of teacher effectiveness on school districts in New York State. This study focuses only on the performance of New York-based teachers, using students' ELA and math assessment outcomes and the New York State APPR ratings for teachers. As such, this study is intended to produce more definitive and reliable findings that will be applicable within the New York State education system and its specific teacher and student population. Ultimately, the need for this study is best illustrated by the conflicting results in the current body of evidence. When considering the substantial inventory of both benefits and detriments associated with teacher effectiveness, it is important to weigh these costs, allowing for the determination of informed decisions. First and foremost, it is a priority to identify the nature of the relationship between teacher effectiveness and student achievement or determine if there is a relationship at all. In the absence of a sufficient, valid relationship between teacher effectiveness ratings and student achievement, further discussion of any pros or cons is useless. It is imperative, then, to further study the relationship between teacher effectiveness and student achievement in an effort to identify more definitive answers.

The methods by which the objectives of this study were achieved are elaborated on in the chapter that follows.

Chapter III

Research Methodology

This chapter presents the research methodology selected for this study and details the various analyses to be applied. The chapter first describes the topic of investigation and briefly presents the aim of this research. These pages also elaborate on the chosen design, as well as the justification for the method at the core of this study. This includes a discussion of the target population and sampling procedures, as well as information relevant to the data. The chapter closes with a description of the analysis to be applied to the data, as well as the potential outcomes and the subsequent insights to be gained.

Relevant Background to the Study

NCLB (2001) ushered in changes that would forever transform the landscape of public education in the United States. In an attempt to ensure equality in American education, the laws required standardization of curriculum, consistent academic standards, and testing systems for the promotion of accountability. These changes led to a continuing focus on comparing the performance of American students in a global capacity and on international tests, specifically. In response to these changes, New York State implemented a series of new requirements for school districts across the state. The subsequent outcome was increased testing and the institution of assessments for the provision of data that would support and promote accountability measures for students, teachers, and principals. In light of these events the relevance of this study is evident and provides necessary insights related to the relationship between teacher effectiveness and student achievement within the state of New York.

Topic and Significance

One of the most prominent issues in the teaching profession today is teachers'

effectiveness. It is potentially a critical element in the success of the students and the overall system of education. Accountability means that everyone is held responsible to high standards of performance. It is paramount to assess the development and learning of students as it helps guide continued growth, effective teaching, and learning. Identification of every student's needs is critical, as it enables educational stakeholders to view learning as a continuum in which student development is noted in different, but equally relevant ways within each student.

This study examined these elements that are critical to the system of education and its success, or lack thereof, as a whole. The aim of this research was to explore and potentially identify the relationship between teacher effectiveness and students' achievement, enabling the provision of recommendations to improve student performance. By understanding the nature of the relationship between teacher effectiveness and student achievement, individual states may be better equipped to direct resources and assistance to the districts and school organizations that are most in need.

Research Design and Methods

The study adopted a quantitative research method. This method involved the collection of quantitative data, analyzing it using statistical and mathematical techniques, and drawing conclusions based on the analysis results (Camerino, Castañer, & Anguera, 2014). The research approach emphasized objective measurement and statistical, numerical, or mathematical analysis of quantitative data. The researcher's specific goal within the context of this non-experimental, correlational, explanatory, cross-sectional quantitative study was to determine the association between teacher effectiveness, the explanatory variable, and student achievement, the response variable. In other words, the independent variable was teacher effectiveness as indicated by average percentage APPR ratings, thereby serving as the independent data set for this study.

Student achievement served as the dependent variable and was represented by student average percentage ELA and math testing scores.

It is important to note that alternative research methods could have been adopted, including qualitative and mixed research methods. The qualitative research methods are designed in a way that assists the researcher in revealing the perceptions of target respondents, typically through open-ended and conversational communication (Yüksel & Yıldırım, 2015). While qualitative methods play a very significant role in research, they are faced with numerous disadvantages, including an inability to quantify relationships or identify a level of significance or cause and effect.

Mixed research methods involve a combination of aspects from qualitative and quantitative research methods. Although the approach might have provided the ability to offset weaknesses inherent in any one methodology, it was not thought to be an optimal fit for the study proposed here. More specifically, according to Bozkurt et al. (2015), the data needed to be transformable in some way to enable application into both types of research approaches, which was not ideal in this study. Also, inequality between the qualitative and quantitative methods could result in unequal evidence within the study, a situation that could be disadvantageous when attempting to interpret the results. Ultimately, the quantitative method was chosen for its ability to incorporate data derived from a large sample that was more representative of the target population and therefore more conducive to extrapolation (Şahin & Levent, 2015). This was complemented by the execution of a quantitative study that allowed for easy replication of procedures and results because of its increased reliability. Ultimately, this meant comparing students' standardized testing average percentage scores with teachers' average percentage APPR ratings in linear regression analyses, ANOVA and the associated models, where

applicable, allowing for the identification of possible relationships and the contribution of one variable to the other (Creswell, 2015).

Target Population

According to the New York State Department of Education, as of the date of this study, there are 62 counties, 732 school districts, 4,782 schools, and 2,622,879 students in New York State. A random sampling was used for this study to collect a smaller sample of the New York State population to make generalizations. Each county was assigned a number and five numbers were selected. The school districts in the five counties were then assigned a number and random.org was utilized to select the numbers assigned to the school districts. Once the school districts were identified, the schools in the district were assigned a number and random.org was utilized to select the schools assigned to be utilized in the study. The sample contained a cross-section of the population of New York State. The schools included were located in urban, suburban, and rural regions of New York State.

The study included elementary and junior high/middle schools within the Orange County, Wyoming County, Westchester County, Nassau County, and Suffolk County regions in New York State. This translates into a total of 37 school districts, including 155 schools that were of relevance from within these respective districts. When looking at enrollment data and teacher employment for each of the schools included, the size of the student and teacher population examined totaled 93,340 students and 6,915 educators. Table 3.1 illustrates each of the aforementioned counties, the specific school districts within each, the number of schools that qualified for inclusion within each of these districts, as well as the number of students and teachers for each individual school district. This is supplemented with information reporting county totals for the number of schools, students, and teachers that made up the data for each

county. This depicts the distribution of schools across the various counties, as well as the proportion of participants derived from each county that compiled the sample as a whole.

Table 3.1

Student and Teacher Sample Population Totals

County School District	No. of Schools	No. of Students	No. of Teachers
Orange County			
Port Jervis City SD	3	1,781	116
Greenwood Lake UFSD	2	529	49
Pine Bush CSD	6	3,362	259
Newburgh City SD	11	7,643	638
Chester UFSD	2	1,068	83
Florida UFSD	1	365	32
Tuxedo UFSD	1	131	12
Cornwall CSD	4	2,102	144
Middletown CSD	5	5,141	373
Orange County Totals	35	22,122	1,706
Wyoming County			
Attica CSD	2	825	75
Perry CSD	2	783	76
Letchworth CSD	2	618	54
Warsaw CSD	2	856	90
Wyoming CSD	1	114	17
Wyoming County Totals	9	3,196	312
Westchester County			
Yorktown CSD	4	2,231	175
Katonah-Lewisboro UFSD	4	2,037	168
Byram Hills CSD	2	1,106	99
Mt. Vernon SD	12	5,121	404
Lakeland CSD	6	3,759	279
Ossining UFSD	3	2,208	77
Scarsdale UFSD	6	3,273	259
Porter Chester-Rye UFSD	5	3,314	226
Greenburgh CSD	3	954	87
New Rochelle City SD	8	7,110	498
UFSD Tarrytowns	2	1,238	89
Bedford CSD	6	2,855	232
Westchester County Totals	61	35,206	2,593

Table 3.1 continued

County School District	No. of Schools	No. of Students	No. of Teachers
Westbury USFD	4	3,082	197
Herricks UFSD	4	2,554	209
Malverne UFSD	2	768	71
Garden City UFSD	3	2,150	155
Uniondale UFSD	7	4,481	396
Nassau County Totals	20	13,035	1,028
Suffolk County			
Sayville UFSD	4	2,012	147
Southold UFSD	2	783	83
Amagansett UFSD	1	93	22
Middle Country CSD	10	5,649	378
Springs UFSD	1	713	65
Brentwood UFSD	12	10,531	581
Suffolk County Totals	30	19,781	1,276
Totals All Counties	155	93,340	6,915

Note. Data collected and aggregated from data.nysed.gov

Instruments

Within the context of this study, the instruments utilized for measuring the variables of teacher effectiveness and student achievement were the standardized tests of proficiency and performance administered in the academic environment. However, unlike other studies, the secondary nature of the data used means that the instruments were previously administered for assessment and measurement of these variables, thereby negating the need for this researcher to administer any evaluative instruments or tools for assessment. As a result, typical concerns related to appropriate administration for the mitigation of bias or issues of validity and reliability had already been addressed by the New York State Department of Education (NYSED).

APPR Ratings as an Evaluative Tool for Teacher Effectiveness

The APPR is the instrument used for testing teacher effectiveness in the state of New York and, as such, was the evaluative tool for measuring teacher effectiveness within this study.

All data related to this assessment was derived from the 2015–2016 reported results comprised of 606 districts, BOCES, and charter schools, which operated under Education Law §3012-c with an approved Hardship Waiver (Keddie, 2015). Also in school year 2015–2016 student achievement scores were not allowed to be used as a factor in computing teaching effectiveness scores (APPR).

According to the NYSED (2019) in assessing a teacher's performance, a final, overall composite score is calculated for each teacher, which is comprised of various components. Although there may be some subjectivity in implementation or grading criteria that varies by school district, there are three primary areas of assessment, including observation of a teacher's performance in the classroom, student growth, and student achievement. The observation element consists of 60% of the composite score and is based upon New York State Teaching Standards. Student growth and student achievement each provide 20% of the final score. Student growth is represented by student learning across the academic year, while student achievement measurements varies by district. The total of these scores is summed on a scale of 1 to 100 and then transformed into a composite score. However, during the school year 2015–2016 the Board of Regents in New York State along with Governor Cuomo issued a moratorium on the use of student test scores. A taskforce was formed to study the effects of Common Core (nysed.gov). In the end, teachers were rated as 1 = *Ineffective*, 2 = *Developing*, 3 = *Effective*, and 4 = *Highly Effective* (NYSED, 2019). According to one study, effective teachers are likely to provide student-related results that have a lower measure of variation among the students (Sloat, Amrein-Beardsley, Tenpe, & Sabo, 2018). This study used the end-of-year exam results teacher APPR scores from New York State that included principal and superintendent observations of teachers.

New York State ELA and Math Assessments

Perullo and Princeton (2003) observed that it is mandatory for all students in Grades 3–8 in New York State to take the ELA and math tests. The test is given over three days in either January or February. The ELA test encompasses one listening selection and several reading selections. Perullo and Princeton (2003) further stated that students are asked several short answer items, as well as extended response questions, in addition to 28 multiple-choice questions.

After the marking of the test, performance is reported as a scale score and in relation to the performance level. The number of points a student earns is converted to a scale. These scale scores are then used to compare student achievement from one grade to another, as well as from year to year. In terms of performance, scale scores are categorized into four categories, with each category representing one performance level: level 1 represents *not proficient*, level 2 means *partially proficient*, level 3 indicates a score that is *proficient*, and level 4 indicates the performance is *advanced* (Perrullo & Princeton, 2003). The system only considers students in level 3 and 4 to have attained the set ELA and math standards. Perullo and Princeton (2003) pointed out that teachers use scale scores to determine student promotion, placement, and special program decisions. Also, these scores are used to determine which students need tutoring, remedial services, or summer school.

McCombs, Kirby, and Mariano (2010) posited that New York State developed the ELA and math tests in response to NCLB demands. As such, this assessment replaced the previous spring assessments administered to students in Grades 3 through 7 in two subjects only. This test is another product of the standardization and teacher effectiveness movement (McCombs, Kirby, & Mariano, 2010).

ELA and Math Standardized Tests

The variable of student achievement in this study was measured using the New York State ELA and math standardized tests. Similar to the APPR, results are measured on a scale from 1 to 4 with higher scores indicating better performance (NYSED, 2019). Typically, level 1 is indicative of performance that is below grade level, level 2 is identified as representing student performance that is *partially proficient*, but not up to the expected level related to common core standards for the grade, level 3 refers to *proficient* performance, while level 4 indicates a student is *highly proficient* (NYSED, 2019).

In this study, student achievement was measured using the 2015–2016 third through eighth grade New York State ELA and math state test results. In terms of scoring accuracy and credibility, the literature reports that the data was compiled with the help of scoring materials used by scoring leaders who trained the educators how to correctly score the constructed-response questions (Ronfeldt, Farmer, McQueen, & Grissom, 2015). The files included scoring rubrics and a sample student response for each score point that could be attained. Further, annotations were made available with sample responses to help illustrate how scores were obtained (Egalite, Kisida, & Winters, 2015).

Data Collection

Data from the NYSED were used to access and collect APPR ratings for teacher effectiveness, as well as ELA and math outcomes representing student achievement. All data corresponded to the 2015–2016 academic year. Student achievement data included ELA and math results, as well as further categorization of results by county, district, and classification related to students who qualified for free or reduced lunch as a means of assessing socioeconomic status.

Data Analysis

The researcher used Microsoft Excel as a means of initially compiling the data. This data was then transferred to SPSS version 25 software for further analysis. This allowed the ability to screen the data in terms of missing values or outliers. Although incorrect values needed to be manually identified by visually scanning the data, the software had the ability to identify and account for missing data or outliers, thereby making this preferable to the original Excel format. This was important to deter possible issues of bias stemming from missing data points, as well as subsequent limitations resulting from fewer data points for analysis (Camerino et al., 2014; Creswell, 2015). Outliers were excluded because of the potential for skewed results and misleading conclusions emerging as a function of this possibility (Camerino et al., 2014).

Descriptive Outcomes

The analysis involved the computation of descriptive statistics. Tables and charts were used when applicable for the presentation of participant data and comparison, which included average percentage APPR ratings for teachers as well as average percentage ELA and math outcomes for students. In each case, the standard deviation range, as well as minimum and maximum values were reported.

The Relationship Between APPR Ratings and ELA and Math Scores

Inferences regarding the association between teacher effectiveness and student achievement was explored using correlation analysis for identifying if a possible relationship existed. Based on the work of Forman & Markson (2017) on possible underreporting of *effective* and *ineffective* teachers, and analysis of teacher ratings of *highly effective* and *effective* ratings in relationship to student achievement ratings was conducted. Specifically, a Pearson's correlation coefficient was the product of the analysis between APPR ratings as a measure of teacher

effectiveness, and ELA and math results as a measure of student achievement. Although this does not define a cause and effect relationship, correlation is a method of statistical evaluation that researchers use to study the strength of a relationship between two, numerically measured continuous variables (Cohen, West, & Aiken, 2014). This analysis provided the ability to determine if a relationship exists between these variables, as well as the strength and direction of any relationship identified (Cohen et al., 2014). This served as a method of preliminary analysis.

Hierarchical Linear Regression

The aforementioned correlation analysis was then further examined through the application of a hierarchical linear regression analyses. This allowed for added insights at the school level, exploring the influence and subsequent variations from several potentially influential factors. Overall, ultimately this identified if teacher effectiveness, as indicated by APPR ratings, has an overall bearing on student achievement, while accounting for additional variables. This entailed comparing several models in which each model built upon the previous framework, adding layers of variables (Cohen et al., 2014).

All models focused on the APPR ratings as an indicator of teacher effectiveness and student achievement as the dependent variable, as indicated by average percentage ELA and math scores. The first model included student factors such as the average percentage of lower socioeconomic status. The second model for comparison controlled for school profile factors, such as the average percentage of students receiving free or reduced lunch and average class size. The third model involved teacher variables, including the influence of teachers with a master's degree or higher and experience. The final model included the average percentage APPR ratings, allowing for the impact of this variable to be evaluated above all others.

The end objective was not only to discover the relationship between the two primary

variables of interest, but also to identify the degree of variance in the dependent variable, student achievement, that was explained within each model. This allowed for greater insights into the impact of APPR ratings while controlling and considering the impact of additional variables.

Hence, the researcher employed a hierarchal linear regression analysis for further evaluating the possibility that variations in teacher effectiveness might trigger changes in student achievement (Cohen et al., 2014; Haghighat, Abdel-Mottaleb, & Alhalabi, 2016). In addition, the use of ANOVA within these models allowed for the identification of changes in R_2 between each model and the extent to which variations in student performance are a product of APPR ratings, or vice versa, as indicated by the corresponding p values (Cohen et al., 2014). The results chapter provides tables of all coefficients and changes between models.

Limitations of the Study

The study faced several limitations. The first one was related to the sampling method adopted. Compared to the simple random sample, the stratified sampling technique required more administrative efforts and the analysis was computationally more complex (Yüksel, & Yıldırım, 2015).

Also, the study used a linear regression model to assess the effect of teacher effectiveness on student achievement. These models can only explain variations in the response variable that can be attributed to variations in the explanatory variables applied (Bozkurt et al., 2015). However, according to the information available in the literature, many variables may influence the variations of student achievement including support and availability of parents, the geographical location of the education institution, the diversity of student profiles, etc. Hence, this study only accounts for the effect of teacher effectiveness as demonstrated through the variables associated with each model, which fail to account for the effects of other factors that

may influence the variations in student achievement. All results are reported in the findings of the final study, accompanied by a discussion of the results and the insights gained from their interpretation.

Chapter IV

Results

The purpose of this quantitative study was to examine the potential link between teacher effectiveness in New York State and its possible relationship to student achievement. Two goals emerged from the question and are compatible with the purpose: (a) explore the relationships among the student factors, teacher characteristics, school factors, teacher APPR ratings, and student achievement on New York State ELA tests at the school level; and (b) explore the relationships among the student factors, teacher characteristics, school factors, teacher APPR ratings, and student achievement on New York State math tests at the school level. The study was motivated by the following research questions:

I. What is the relationship between teacher effectiveness and achievement in ELA and math at the school level when controlling for student characteristics (enrollment, free and reduced lunch, and economically disadvantaged)?

A. ELA with controls

B. Math with controls

II. What is the relationship between teacher effectiveness and student achievement in ELA and math at the school level when controlling for teacher qualifications (experience and highest degree)?

A. ELA with controls

B. Math with controls

III. What is the relationship between student achievement in ELA and math and teacher effectiveness (APPR ratings) at the school level?

A. ELA without controls

B. Math without controls

The methodology used in this quantitative correlational design consisted of correlation and hierarchical linear regression modeling. Variables were defined and operationalized for the study. The effectiveness of teaching was operationalized as APPR ratings at the school level, while the dependent variable, student achievement, was evaluated as student performance on New York State ELA and math tests. These test scores consisted of the average percentage of students scoring at ELA and math standards at the school level and as defined by level 4 (*highly proficient in standards*), level 3 (*proficient in standards*), level 2 (*partially proficient in standards*), and level 1 (*well below proficient in standards*). The variables for teacher effectiveness (APPR scores) and student achievement (NYS ELA and math scores) were used in both the correlation and regression analysis.

The student characteristics were gender, disability status, and economic status. School profile factors were enrollment, average class size, and free or reduced lunch. Teacher factors or characteristics were defined as the average percentage of those who held master's degrees, doctoral degrees, and fewer than three years of experience.

For each of the goals null and alternative hypotheses were formulated and tested:

H1o: Student characteristics, teacher characteristics, and school characteristics at the school level do not jointly and significantly predict student achievement defined by ELA scores.

H1a: Student characteristics, teacher characteristics, and school characteristics at the school level jointly and significantly predict student achievement defined by ELA scores.

H2o: Student characteristics, teacher characteristics, and school characteristics at the school level do not jointly and significantly predict student achievement defined by standardized math scores.

H2a: Student characteristics, teacher characteristics, and school characteristics at the school level jointly and significantly predict student achievement defined by standardized math scores.

H3o: Teacher APPR scores at the school level do not jointly and significantly predict student achievement defined by standardized ELA and math scores.

H3a: Teacher APPR scores at the school level jointly and significantly predict student achievement defined by standardized ELA and math scores.

The results are organized as descriptive, correlation analysis, followed by the results for hierarchal regression model tested. The chapter concludes with a summary.

Demographics

This study targeted students in Grades 3 through 8 in New York State and their performance on the 2015–2016 New York State ELA and math tests. The study sought to use a cross-sectional population of students. The student demographic data for the counties included in the study are included in Table 4.1.

Table 4.1

Student County Sample Demographics 2015–2016

County	Totals	Avg. Percent
Orange County		
Male Students	1,801	51.45
Female Students	1,689	48.25
American Indian/Alaska Native	5	0.14
Black Students	522	14.91
Hispanic Students	1,147	32.77
Asian, Native HI, Pac. Island	98	2.80
White Students	1,609	45.97
Multiracial Students	118	3.37
Students with Disabilities	544	21.76
Economically Disadvantaged	1,628	46.51
Wyoming County		
Male Students	453	50.33
Female Students	1447	49.67

Table 4.1 continued

Wyoming County	**Totals**	**Avg. Percent**
American Indian/Alaska Native	2	0.22
Black Students	6	0.66
Hispanic Students	21	2.33
Asian, Native HI, Pac. Island	6	0.66
White Students	849	94.30
Multiracial Students	14	1.55
Students with Disabilities	111	12.33
Economically Disadvantaged	410	45.55
Westchester County		
Male Students	3,127	51.26
Female Students	2,973	48.73
American Indian/Alaska Native	5	0.08
Black Students	1,289	21.13
Hispanic Students	1,833	30.05
Asian, Native HI, Pac. Island	333	5.46
White Students	2,504	41.05
Multiracial Students	131	2.15
Students with Disabilities	864	14.16
Economically Disadvantaged	2,461	40.34
Nassau County		
Male Students	1,038	52
Female Students	962	48
American Indian/Alaska Native	0	0
Black Students	476	24
Hispanic Students	774	39
Asian, Native HI, Pac. Island	282	14
White Students	444	22
Multiracial Students	20	1
Students with Disabilities	259	13
Economically Disadvantaged	1,081	54
Suffolk County		
Male Students	1,495	49.83
Female Students	1,453	48.43
American Indian/Alaska Native	1	0.03
Black Students	158	5.27
Hispanic Students	1,344	45
Asian, Native HI, Pac. Island	102	3.4
White Students	1,350	45
Multiracial Students	49	1.63
Students with Disabilities	444	15
Economically Disadvantaged	1,560	52

Note. Data collected and aggregated from data.nysed.gov

The study also focused on teachers in New York State for the school year 2015–2016 and their APPR data percentages and averages at the school level. During this school year the student achievement scores (NYS ELA and math) were not allowed to be used as a factor to compute the teacher effectiveness scores (APPR). During the school year 2015–2016, there were 210,496 teachers in New York State. Eight percent of teachers had fewer than three years of experience and 39% held master's degrees plus thirty hours or doctorates. The five counties included in the research had an average percentage of 4.14 percent of teachers with fewer than three years of experience. A total of 6,915 teachers were included in the study, with an average percentage of 33% of teachers with master's degrees plus thirty or doctorates, noted in Table 4.2.

Table 4.2

County Teacher Sample Demographics 2015–2016

County	Avg % MS+/Doctorate	Avg % Fewer than 3 Years	Total Teachers
Orange County	29.5	4.5	1,706
Wyoming County	10.4	5.9	312
Westchester County	55.2	2.9	2,593
Nassau County	60.9	4.0	1,028
Suffolk County	79	3.4	1,276
Totals	33	4.1	6,915

Note. Data collected and aggregated from nysed.gov

Aggregate Outcomes for ELA and Math

The school and district data were aggregated to allow overview and background for the results. The specific school districts and associated counties included in this study were detailed in the previous chapter. These aggregate student data for the schools are represented as average percentages of ELA and math standardized test outcomes across all students and schools.

ELA Performance

The ELA was scored using categories that indicate the achievement levels students have attained relative to the expected for grade level**.** The ELA test data were aggregated as the

percentage of students who performed at each category or level. The possible values used for ratings were integers between 1 and 4 with 1 indicative of the lowest possible score corresponding to the category of *below grade level* and 4 indicating the highest possible score corresponding to *highly proficient* (Table 4.3). The percentage of students scored for each rating was recorded for each school and the mean percentage was calculated for the percentage of schools that achieved each proficiency category score is shown in Table 4.3.

Table 4.3

Aggregated ELA Test Scores Across Schools

Achievement Category	*M* %	SD
1 = Well-Below Proficient	26	13.44
2 = Partially Proficient	34	7.27
3 = Proficient	29	10.79
4 = Highly Proficient	11	8.78

Note. Percentages represent the students scoring in each achievement category aggregated across all schools in the sample. $N = 155$, the number of schools in the sample.

From the data in Table 4.3, it appeared more students across all schools scored in the lower achievement categories than in the higher. After combining the two lower achievement categories, 1 with 2, and comparing the result to the two higher combined categories, 3 with 4, the total average percentage scores of students across all schools who scored in the higher achievement categories was less than the average percentage scores of students across all schools that were below proficient (Table 4.4).

Table 4.4

Aggregated ELA Test Scores of Sample Scoring Proficient (versus those that did not).

Rating	*M* %	SD
Scored Proficient or Above (Levels 3 and 4)	40	17.96
Scored Below Proficient (Levels 1 and 2)	60	17.96

Note. Percentages represent the students scoring in the combined ratings aggregated across all schools in the sample. $N = 155$, the number of schools in the sample.

The aggregate results for ELA scores provided insights related to student achievement. The combining of the data into two groups provided a rationale for comparisons in the correlation analysis shown below. Furthermore, student performance was contrasted with additional findings from teacher aggregate APPR ratings across schools in the discussion in Chapter V.

Math Performance Outcomes

These data from math tests were analyzed in the same way as the average ELA student score across all schools. Therefore, the data represent aggregated from the math test scores across all schools were presented as the average percentage of students who performed at each level (Table 4.5). The categories 1 to 4 associated with the student performance are the same as those described above for the ELA results.

Table 4.5

Math Testing Outcomes: Average Percentage of Sample for Each Rating

Achievement Category	*M* %	SD
1 = Well-Below Proficient	27	15.97
2 = Partially Proficient	32	7.99
3 = Proficient	23	8.27
4 = Highly Proficient	18	13.69

Note. Percentages represent the students scoring in each achievement category aggregated across all schools in the sample. $N = 155$, the number of schools in the sample.

In examining the math outcomes, the proportion of students across schools included those who scored *proficient* or *not proficient*. The students' scores at levels 1 or 2, in the *partially* and *well below proficient* categories, equated to an average 58%, while an average 42% of students' scores were in the *proficient* and *highly proficient*, levels 3 or 4, as seen in Table 4.6.

Table 4.6

Aggregated Math Test Scores of Sample Scoring Proficient (versus those that did not).

Rating	*M* %	SD
Scored Proficient or Above (Levels 3 and 4)	58	20.43
Scored Below Proficient (Levels 1 and 2)	42	20.43

Note. Percentages represent the students scoring in the combined ratings aggregated across all schools in the sample. *N* = 155, the number of schools in the sample.

Aggregated Teacher APPR Ratings from Sample

The teacher ratings are reported in aggregate, as were the student achievement data. The ratings categories for teacher APPR are from 1 as the lowest rating category, and up to 4, indicating the highest category. Like the student achievement data, the APPR data are reported as the average percentage of teachers in each category across all schools (Table 4.7).

Table 4.7

Teacher APPR Outcomes: Average Percentage Across Schools for Each Rating

Achievement Category	*M* %	SD
1 = Well-Below Proficient	.50	1.76
2 = Partially Proficient	2.50	5.48
3 = Proficient	41	29.57
4 = Highly Proficient	56	31.29

Note. Percentages of teachers aggregated across school in each rating category aggregated. *N* = 155.

The substantial majority of teacher participants' ratings across schools were among the higher APPR ratings, indicating that most teachers were either *effective* or *highly effective* (Table 4.8). This contrasted with the student achievement data by category at school level, which showed the student scores appeared to be more widely distributed across the performance categories. The next step entailed categorizing the teacher APPR ratings into two subgroups: teachers categorized as *effective*, indicated by APPR ratings 3 or 4, and teachers who were not,

indicated by APPR ratings of 1 or 2. The average percentage of teachers whose rating were grouped in the two lower performance categories was almost negligible, $M = 3\%$. Their more effective counterparts average ratings across schools were among the higher ratings, $M = 97\%$ (Table 4.8).

Table 4.8

Aggregated Teacher APPR Scores of Sample that were rated "Effective" (versus those that were not).

Rating	*M* %	SD
Scored Proficient or Above (Levels 3 and 4)	97	6.89
Scored Below Proficient (Levels 1 and 2)	3	6.89

Note. Percentages represent the teachers rating in the combined ratings aggregated across all schools in the sample. $N = 155$, the number of schools in the sample.

As indicated in Table 4.8 most teachers were categorized in the higher performing proportion of the sample. This sharply contrasted with the student achievement data in which the larger average percentage of students across the schools fell into the lower performing categories. The relationships between the subgroups of aggregate teacher ratings and student scores across all schools were examined using correlation analysis.

Correlation Analysis

For assessing the viability of regression modeling for the data, correlation analysis was used to make inferences regarding the relationships among the primary variables. Pearson's *r* was used to test the association between the average percentage of teacher APPR ratings aggregated across schools and student achievement as measured by the average percentage ELA and math results aggregated across schools. If the correlation was significant, it would provide evidence of strength and direction of a relationship between these primary variables (Cohen et al., 2014).

These data for students and teachers across schools were regrouped such that the two higher levels of performance were combined and similarly the two lower levels of performance were grouped. The rationale behind this was straightforward: If student performance was significantly correlated with the performance of teachers, then there would be strong correlations between average percentages of higher-performing students and average percentages of higher-performing teachers across all schools. These tests allowed for easier identification of potential relationships that might occur in the higher performing groups regarding teacher effectiveness and student achievement.

Bivariate Correlation Analysis: APPR and ELA

The correlation analysis concerned the relationship between average percentage APPR ratings and average percentage student ELA standardized test scores. The data were computed as average percentages, and therefore, could be considered as continuous variables suitable for this analysis. A Pearson *r* was computed to test the relationship between the average percentage of higher performing students and higher performing teachers.

The average percentage of teachers with higher APPR ratings and the average percentage of students with higher ELA test performance were positively correlated, as indicated by, $r\,(155) = .33$, $p = .000$. Thus, as the average percentage of teachers with higher APPR ratings increased, the average percentage of higher performing students on the ELA test increased. The coefficient value of .33 indicated that the size of this relationship was small and nine percent of the variation in the percentage of students proficient in ELA could be explained by the percentage of teachers scored as effective.

Bivariate Correlation Analysis: APPR and Math

The correlation between higher student math score average percentages, 3 or 4, and the

higher teachers' APPR average percentage ratings of, 3 or 4, was tested. The underlying assumptions were generally those described above for the correlations between ELA scores for average percentage of students across schools and teacher APPR average percentage ratings across all schools. It was expected that average percentages of higher performing teachers would be positively correlated with the average percentages of higher performing students on math scores. There was a positive correlation between the average percentage of teachers with higher APPR ratings and the average percentage of students with higher math test scores. The Pearson coefficient was significant, r (155) = .34, p =.000. As in the ELA results, the coefficient of .34 indicated a positive, yet small relationship size and nine percent of the variation in the percentage of students proficient in math was explained by the percentage of teachers scored as effective.

Comparisons of Results from ELA and Math Correlations with APPR Ratings

There was consistency among the correlations for ELA average percentage scores and teacher average percentage APPR ratings, and those for math test average percentage scores and teacher average percentage APPR ratings. The relationship between the average percentages of teachers rated higher on their APPR and the average percentage of students performing higher on the standardized tests was reflected by positive associations for both tests.

The pattern of results was similar for the two dependent variables ELA testing outcomes and math test outcomes when considering the higher scoring student subgroup and the higher rated teacher subgroup. The results showed a statistically significant positive correlation between the average percentage higher APPR-rated teachers and the average percentage of higher-scoring students on ELA and math standardized tests, although small in both cases.

Hierarchal Linear Regression Modeling

The aforementioned correlations were then further examined through the application of

hierarchal linear regression. This allowed for added insights at the school level, exploring the influence and subsequent variations from several potentially influential factors. In this regard, it allowed the ability to control for extraneous or confounding variables, ultimately providing more accurate information pertaining to whether or not teacher effectiveness, as indicated by APPR ratings, had a relationship to student achievement. In addition, this also provided added insight into whether APPR ratings were a sufficient indicator of teacher effectiveness, in and of themselves. More specifically, hierarchal linear regression provided an optimal method for evaluating the relationship of APPR ratings, or teacher effectiveness, on the dependent variable, student achievement, by including all other potentially influential variables first and then adding in the variable of APPR ratings in the final step. By adding the variable of teacher effectiveness, as indicated by APPR ratings, last, after consideration of other variables, it was possible to identify the proportion of the dependent variable explained by this factor, while also observing how much this may have changed from the prior models (Cohen et al., 2014). The APPR ratings and its potential relationship to student achievement was examined using linear regression analysis.

APPR Ratings Relationship to Student Achievement

The hierarchal linear regression models included the average percentage APPR ratings and the relationship of this variable on student achievement, as indicated by average percentage ELA and math standardized testing outcomes. The APPR ratings were included cumulatively, in addition to all of the aforementioned variables. This variable, as the independent variable of interest, was added in order to explore the relationship of this factor above and beyond all other potentially influential variables. In doing so, it was possible to observe how much this changed because of the influence of APPR ratings alone.

As previously discussed, the average percentage of teachers who received an APPR rating of 3 or 4 was added for an overall average percentage of teachers at each school who were rated as *effective* or higher. Once again, the assertion was that the average percentage of *effective* teachers would influence the average percentage of *proficient* students. Therefore, as the average percentage of *effective* teachers increased, so would the average percentage of *proficient* students. This model was first run with the average percentage of *effective*, or a rating of 3, and *highly effective*, or a rating of 4, teachers combined for an overall average percentage of effective teachers at each school.

This analysis was then repeated, utilizing only the average percentage of *highly effective* teachers at each school—only those who scored a 4 regarding their APPR rating. The underlying purpose was to examine only the association of the highest rated teachers to explore how much of an influence was realized when comparing the influence of the most effective educators, as opposed to the inclusion of *effective* teachers and higher. This was also motivated by the assertion in the prior research that an APPR rating of *effective* might be the new *ineffective* (Forman & Markson, 2015) and, therefore, if *effective* teachers, scoring a 3 on APPR ratings, were actually not effective, it would impede the accuracy of results and the subsequent influence on student achievement. Thus, by including only the educators who were rated the highest and were considered effective, regardless of the authenticity of other ratings, then a more accurate assessment of the influence on student achievement would be possible. For this reason, a separate analysis was executed using only the average percentage of educators who received an APPR rating of 4 and student achievement scores of proficient.

The objective of this methodology was to determine the relationship between the two primary variables of interest and also to identify the amount of variance in the dependent

variable student achievement. The order of the variables was chosen because of its ability to illustrate the association of APPR ratings, while controlling and considering the influence of additional variables. The overarching hypothesis was that variations in teacher effectiveness, as indicated by APPR ratings, would trigger changes in student achievement (Cohen et al., 2014; Haghighat, et al., 2016). In addition, the use of ANOVA allowed for the identification of changes in R_2 between each model and the extent to which variations in student performance were a product of APPR ratings, or vice versa, as indicated by the corresponding p-values (Cohen et al., 2014).

ELA Outcomes

The hierarchal linear regression involved the average percentage of students who scored 3 or 4 on the ELA standardized tests. This served as the variable of student achievement. The teacher scores of effective and highly effective served as the variable of teacher effectiveness. Table 4.9 shows the model Summary Output derived from SPSS.

Table 4.9

Model Coefficients and Summary for Average Percentage APPR Ratings ("Effective+" 4 rating) and ELA Test Outcomes

	Model 1	Model 2	Model 3	Model 4
% Disabled Test-Takers	-0.46***	-0.28**	-0.28	-0.27
	(0.15)	(0.14)	(0.06)	(0.06)
% Economically Disadvantaged	-0.43***	-0.01	-0.01	-0.01
	(0.03)	(0.07)	(0.93)	(0.93)
% Females	-0.02	-0.00	0.00	0.00
	(0.06)	(0.05)	(0.98)	(0.98)
Enrollment		-0.00	-0.00	-0.00
		(0.40)	(0.40)	(0.44)
Average Class Size		0.77**	0.77**	0.77**
		(0.31)	(0.02)	(0.03)
% Free Lunch		-0.51***	-0.51***	-0.51***
		(0.08)	(0.00)	(0.00)
% Reduced Lunch		-0.51**	-0.51**	-0.52***
		(0.22)	(0.03)	(0.03)
% Teaching Fewer than 3 Years			-0.16	-0.16
			(0.41)	(0.41)
% Teachers w/ Masters/Doctoral			-0.01	-0.00
			(0.90)	(0.92)
Avg. % APPR Ratings Effective +				0.00
				(0.91)
[Avg % 3's and 4's]	66.243	52.899	53.144	52.869
	(3.627)	(7.363)	(7.403)	(7.778)
N	155	155	155	155
R^2	0.67	0.76	0.76	0.76
F of R^2 change	101.12	13.05	0.34	0.01

*Note. * p<0.10; ** p<0.05; *** p<0.01. Standard errors are shown in parentheses. The dependent variable is the percentage of students scoring proficient or highly proficient in ELA.*

As indicated by R^2, each of the models, 1 through 4, accounted for a greater amount of the variance in the dependent variable than the prior model. With the first model accounting for an average 67% of the variation in student achievement, producing $R^2 = .668$, F (3,151) = 101.12, p=.000, and the second model accounting for an average 76%, producing $R^2 = .755$, F (4,147) = 13.05, p=.000, this was a substantial increase. However, model 3 accounted for 76%,

producing $R^2 = .756$, F (2, 145) = .34, p=.713, demonstrating no increase from the prior model. While the addition of APPR ratings in model 4 accounted for an average 76%, producing $R^2 = .756$, F (1, 144) = .01, p=.905 of the variation in student achievement demonstrates no notable improvement from the prior model.

As indicated in Table 4.9, in the final model, only the average percentage of students receiving free lunch, the average percentage of reduced lunch, and class size were statistically significant predictors of student achievement, as indicated by the average percentage of students who scored levels 3 and 4 on the ELA exam. The average percentage of disabled test-takers proved to have a statistically significant negative association to student achievement in the models 1 and 2. The average percentage of disabled test-takers also proved to have a negative association with student achievement in models 3 and 4 but not at a statistically significant level.

ELA Prediction with Only Highly Effective APPR Ratings

The second hierarchal linear regression involved the average percentage of students who scored 3 or 4 on the ELA standardized tests as the variable of student achievement. However, APPR ratings were tested using only the average percentage of teachers with scores of 4 or rated as *highly effective*. Table 4.10 shows the Model Summary and Coefficients Output derived from SPSS.

Table 4.10
Model Summary and Coefficients for Average Percentage APPR Ratings ("Highly Effective" 4 rating) and ELA Test Outcomes

	Model 1	Model 2	Model 3	Model 4
% Disabled Test-Takers	-1.08***	-0.30	-0.29	-0.29
	(0.00)	(0.08)	(0.10)	(0.10)
% Economically Disadvantaged	0.27***	-0.02	-0.02	-0.02
	(0.00)	(0.59)	(0.61)	(0.59)
% Females	0.37	0.22	0.23	0.23
	(0.77)	(0.71)	(0.07)	(0.07)
Enrollment		-0.00	-0.00	-0.00
		(0.88)	(0.86)	(0.88)
Average Class Size		1.06**	0.99***	0.98***
		(0.01)	(0.02)	(0.02)
% Free Lunch		-0.59***	-0.59***	-0.59***
		(0.00)	(0.00)	(0.00)
% Reduced Lunch		-0.49**	-0.42***	-0.42***
		(0.09)	(0.13)	(0.13)
% Teaching Fewer than 3 Years			0.02	0.01
			(0.94)	(0.96)
% Teachers w/ Masters/Doctoral			0.02	0.02
			(0.65)	(0.64)
Avg. % APPR Ratings Effective +				0.01
				(0.86)
[ELA avg. % 3's and 4's]	27.660	39.996	39.772	39.2299
	(8.654)	(10.192)	(10.266)	(10.776)
N	155	155	155	155
R^2	0.31	0.70	0.70	0.70
F of R^2 change	23.00	47.51	0.11	0.30

*Note. * p<0.10; ** p<0.05; *** p<0.01. Standard errors are shown in parentheses. The dependent variable is percentage of students scoring proficient or highly proficient in ELA.*

As indicated by R^2, each of the models, 1 through 4, accounted for no increased amount as compared to the prior model. However, the addition of APPR ratings, including only those that are a 4, or highly effective, produced a small increase from model 3 at an average 76% producing R^2 =.756, F (2,145)=0.34, p=.713 to model 4 at an average 76%, producing R^2 =.756, F (1,144)=0.20, p=.654. Table 4.10 illustrates the associated coefficients output for models 1 through 4. The analysis indicated that average percentage of economically disadvantaged, the

average percentage of free lunch, and average percentage of disabled test-takers were predictors that were statistically significant.

Math Outcomes

This analysis involved the average percentage of students who scored a level 3 or 4 on math standardized tests. This served as the variable of student achievement and the dependent variable in all models. Table 4.11 shows the Model Summary Output derived from SPSS.

Table 4.11

Model Summary and Coefficients for average % APPR Ratings (3 or 4) and Math Test Outcomes

	Model 1	Model 2	Model 3	Model 4
% Disabled Test-Takers	-1.08***	-0.30	-0.29	-0.29
	(0.00)	(0.08)	(0.10)	(0.10)
% Economically Disadvantaged	0.27***	-0.02	-0.02	-0.02
	(0.00)	(0.59)	(0.61)	(0.59)
% Females	0.37	0.22	0.23	0.23
	(0.77)	(0.71)	(0.07)	(0.07)
Enrollment		-0.00	-0.00	-0.00
		(0.88)	(0.86)	(0.88)
Average Class Size		1.06***	0.99****	0.98***
		(0.01)	(0.02)	(0.02)
% Free Lunch		-0.59***	-0.59***	-0.59***
		(0.00)	(0.00)	(0.00)
% Reduced Lunch		-0.49**	-0.42**	-0.42***
		(0.09)	(0.13)	(0.13)
% Teaching Fewer than 3 Years			0.02	0.01
			(0.94)	(0.96)
% Teachers w/ Masters/Doctoral			0.02	0.02
			(0.65)	(0.64)
Avg. % APPR Ratings Effective +				0.01
				(0.86)
[Math avg. % 3's and 4's]	27.660	39.996	39.772	39.2299
	(8.654)	(10.192)	(10.266)	(10.776)
N	155	155	155	155
R^2	0.31	0.70	0.70	0.70
F of R^2 change	23.00	47.51	0.11	0.30

*Note. * p<0.10; ** p<0.05; *** p<0.01. Standard errors are shown in parentheses. The dependent variable is the percentage of students scoring proficient or highly proficient in ELA.*

Overall, the independent variables that comprised each model explained the variation in student achievement, to a lesser extent than these variables explained the variation in the average percentage of students scoring level 3 or 4 math scores. Further, while there was a substantial difference between model 1 and model 2, related to the amount of variance explained by factors included in each (31%/ *r* square = .314 versus 70%/ *r* square = .701, respectively), the added proportion of variation explained in models 3 and 4 was negligible, if not nonexistent.

In Table 4.11 model 4 indicated that the average class size and the average percentage of students receiving free lunch were statistically significant predictors of student achievement, as indicated by the average percentage of students who scored a level 3 or 4 on the math tests. In addition, the predictor average percentage of free lunch negatively influenced or reduced student achievement, while the average class size had a positive association. Finally, as the average percentage of *effective* and *highly effective* teachers increased, student achievement also increased, but not to a statistically significant extent.

Math Prediction with Only Highly Effective APPR Ratings

The linear regression was run again using only the average percentage of teachers with scores of 4 or rated as *highly effective*. The student achievement variables were students receiving level 3 and level 4. Table 4.12 shows the Coefficients and Model Outputs derived from SPSS.

Table 4.12

Model Summary and Coefficients for Average Percentage APPR Ratings ("Highly Effective" 4 rating) and Math Test Outcomes

	Model 1	Model 2	Model 3	Model 4
% Disabled Test-Takers	-1.08***	-0.30	-0.29	-0.28
	(0.00)	(0.08)	(0.10)	(0.11)
% Economically Disadvantaged	0.27***	-0.02	-0.02	-0.03
	(0.00)	(0.59)	(0.61)	(0.46)
% Females	0.37	0.22	0.23	0.24
	(0.77)	(0.71)	(0.07)	(0.06)
Enrollment		-0.00	-0.00	0.00
		(0.88)	(0.86)	(0.96)
Average Class Size		1.06***	0.99***	0.96***
		(0.01)	(0.02)	(0.02)
% Free Lunch		-0.59***	-0.59***	-0.60***
		(0.00)	(0.00)	(0.00)
% Reduced Lunch		-0.49	-0.42	-0.43
		(0.09)	(0.13)	(0.12)
% Teaching Fewer than 3 Years			0.02	0.01
			(0.94)	(0.98)
% Teachers w/ Masters/Doctoral			0.02	0.02
			(0.65)	(0.70)
Avg. % APPR Ratings Effective +				0.02
				(0.55)
[Math avg. % 3's and 4's]	27.660	39.996	39.772	39.599
	(8.654)	(10.192)	(10.266)	(10.293)
N	155	155	155	155
R^2	0.31	0.70	0.70	0.70
F of R^2 change	23.00	47.51	0.11	0.36

*Note. * p<0.10; ** p<0.05; *** p<0.01. Standard errors are shown in parentheses. The dependent variable is the percentage of students scoring proficient or highly proficient in ELA.*

Using only those APPR ratings that were a 4, or *highly effective*, produced a minimal improvement from the prior model. There was a substantial difference between model 1 and model 2, related to the amount of variance explained by factors included in each, average 31% versus average 70%, respectively. Model 1 producing $R^2 = .314$, F (3,151) = 23.00, p=.000 and model 2 producing $R^2 = .701$, F (4, 147) = 47.51, p=.000. The added proportions of variation

explained in models 3 and 4 was negligible reporting $R^2 = .701$, $F\ (2,145) = 0.11$, $p = .899$ for model 3 and $R^2 = .702$, $F\ (1, 144) = 0.00$, $p = .548$.

As indicated in Table 4.12, model 4 indicates that only the average percentage of free lunch and average class size were statistically significant predictors of student achievement, as indicated by the average percentage of students who scored a level 3 or 4 on the math tests. In addition, the predictor average percentage free lunch negatively reduced student achievement. Finally, as the average percentage of *highly effective* teachers increased, student achievement also increased, but not to a statistically significant extent.

Chapter Summary

The purpose of the aforementioned analyses and the subsequent outcomes was to explore and identify answers to the research questions that motivated this inquiry. The various analytic components served to provide insights aimed at formulating these objectives. The bivariate correlation identified that relationships do exist between teacher effectiveness and student achievement. Teachers with effective and highly effective ratings positively correlated to students with effective and highly effective tests ratings in both ELA and math.

The regression analysis did not produce relationships with statistically significant influences on students' achievement when controlling for teachers rated effective and highly effective together or highly effective alone, when controlling for other variables. The average percentage of economically disadvantaged, disabled test takers, free lunch, and reduced lunch variables proved to have negative association to student achievement at a statistically significant level. Average class size proved to have a positive association to student achievement in both ELA and math at a statistically significant level. Free lunch was consistently statistically significant in model 4 of both ELA and math analysis that included teacher APPR scores. It also

had a negative association to student achievement.

Teacher effectiveness and student achievement was positively correlated in the bivariate correlational analysis but not when controlled with other variables in the regression analysis. The previously presented results are discussed in detail regarding their interpretation, their implications, their limitations, and recommendations for future study in the final chapter of this dissertation.

Chapter V

Discussion

The objective at the core of this dissertation was to examine the potential link between teacher effectiveness in New York State and its possible relationship to student achievement when measured by standardized test scores These variables were analyzed and the results presented in the previous chapter. The following pages elaborate on the analytic results and subsequent findings, focusing not only on their interpretation, but their implications, possible limitations and, finally, recommendations that were formulated in response to these outcomes.

Interpretation of Results

The preliminary descriptive statistics provided a snapshot of student achievement on standardized tests prior to exploring the influence of other variables. In both cases of standardized test performance, ELA and math results were remarkably similar with only about average 40% ELA and 58% math of those sampled performing at a *proficient* level, scores of 3 or 4. Yet the accompanying pattern of APPR ratings indicated that teacher effectiveness, as a majority average 97%, was *effective* or better, as evidenced by scores of 3 or 4. At first glance, this is intriguing that such predominantly high-scoring teachers would produce such low-performing students. This leads to many pertinent questions, such as how much teachers may actually influence student outcomes, whether APPR ratings serve as a sufficient indicator of teacher effectiveness, as well as whether or not some other factor is responsible for profoundly influencing student performance in New York State.

It was because of these preliminary results that a correlation analysis was executed between APPR ratings and standardized test outcomes. As formerly stated, the hypothesis behind these analyses was that, in general, a greater number of higher rated teachers would result

in a greater number of high-performing students—assuming, of course, that teacher effectiveness influenced student achievement, APPR elements should predict student achievement, *and* that APPR ratings were an accurate indicator of teacher effectiveness. When looking at the higher performing teachers, in terms of average percentages, and the associated higher performing student average percentages for ELA and math outcomes, in both cases, the resulting correlations were positively correlated and to a statistically significant extent. In other words, as the average percentage of high-performing teachers increased, the average percentage of higher performing students increased as well.

The positive correlations presenting between APPR ratings and student outcomes (both ELA [r (155) = .33, p = .000] and math tests [r (155) = .34, p =.000]) indicated that the greater the average percentage of *higher performing* teachers at each school, the greater higher-performing students, in average percentages, were realized, in accordance. In the set of correlations, the more *highly effective* or *effective* teachers there were at the school level, then the more high-performing students would result as a function of these teachers.

Linear Regression Models

The linear regressions were implemented using data from the average number of faculty identified as *effective* or *highly effective*, while also examining the association of those identified as *highly effective* alone. This provided a way of examining how the average number of teachers rated *effective* or better influenced student achievement along with other variables. This also afforded the opportunity to examine the association of *effective* and *highly effective* teachers in a singular manner, providing insight into the influence of each rating alone. This was motivated by the work of Forman and Markson (2015) and their assertions that only *highly effective* teachers were truly effective and those rated as *effective* were simply overrated and were not

genuinely effective.

Thus, if *effective* teachers, those rated with an APPR rating of 3, were actually not effective, this may provide some insight into why such unexpected outcomes occurred when examining the correlations between the average number of teachers rated *effective* or higher, and the related average number of students who scored *proficient* or better on the standardized tests. In either regard, the results were not statistically significant in the subgroup of *highly effective* APPR ratings and the subgroup of combined APPR scores of *effective* and *highly effective* when controlling for other variables.

Further, when looking at the third and fourth linear regression models, these models systematically explored the influence of teacher experience and teacher education, in the first case, followed by the addition of APPR ratings in the final model. This allowed for an assessment of the extent to which APPR ratings may contribute alone, above and beyond all other factors considered. In each case, ELA and math tests, APPR ratings had no statistically significant association to student test outcomes. In fact, the minute change in the model from the prior model configuration was negligible, at best, indicating no improvement in model fit from adding the influence of APPR ratings in relation to student achievement.

Similarly, the same negligible association was realized from model 2 to model 3, reflecting the contribution of teacher effectiveness. In fact, teacher experience—years teaching—had a negative association on ELA outcomes and only a slight positive influence on math outcomes, in which neither was near the mandated criteria for statistical significance. Meanwhile, the average number of teachers with a master's degree or higher had a negative association on both test outcomes. Interestingly, all variables analyzed that represented the quality of teaching had no influence, a negligible association, or a negative association on

student achievement.

In the end, these unexplained findings prompt the question as to whether or not this says something about teachers. Or, perhaps, does it speak to how teacher effectiveness and student achievement is evaluated? Or, ultimately, does it imply the existence of other issues occurring within these schools that influence student achievement in such a widespread manner that it overshadows the influence of teachers altogether?

Factors of Significance

While the previous paragraphs elaborate on a number of output results that detail factors of interest that were not significant, this begs the question of what factors were significant. In some regards, there were no surprises related to one independent variable identified as statistically significant—socioeconomic status. More specifically, in the models analyzed within this study, the average percentage of students receiving free lunch was representative of students who came from a household with a lower socioeconomic status, as this is a qualifying factor in free lunch eligibility. Similarly, the average percentage of students receiving reduced lunch at each school also served as an indicator of socioeconomic status (SES), but not to the extent of free lunch student subgroups.

Nevertheless, one may say that the role of SES as a predictor of student achievement was predictable in and of itself. In fact, the negative influence of SES on student achievement is a repetitive theme and a frequently recognized finding in many prior studies, such as that of Berliner (2013, 2014). The results of this study reaffirm the aforementioned findings, as well as the findings of many other research endeavors that have realized the same results.

Another variable found within the context of this study was the number of disabled test-takers. Or more specifically, the higher the average number of students with disabilities at the

school level, the higher the average percentage of low-performing students on both ELA and math test results. While disabled test takers had a negative association on student achievement throughout the models it was not always at the level of statistical significance.

Finally, average percentage class size was an influencing factor across models, in both ELA and math test outcomes. Also differing from the prior variables, class size actually had a positive association on student test outcomes, while a growing number of students with disabilities or a greater average number of lower income students produced a negative association. In terms of average class size, as class size increased, student performance on the ELA and math tests increased. The assumption proposed that the larger the class, the less individual attention each student receives and the likely result would be lower achieving students. Yet the opposite occurred in this case.

These findings replicate those of Berliner (2013, 2014) in which students of a higher social class were associated with increased proportions of students who passed, while conversely, students receiving free lunch were associated with higher fail rates (Berliner, 2013). However, the factors that were found to present with a significant influence on student achievement were variables that had a negative association. These variables are also not within the schools control. Therefore, this offers little insight into how to promote, improve, or increase student achievement. Conversely, focusing on these students to improve these subgroup testing scores may serve to somewhat improve student achievement. Perhaps, New York State should take a look at how these students are tested; for example, students with disabilities are functioning at minimum two grade levels below their assigned grade, yet they are assessed using the test for their assigned grade level as opposed to testing them on the academic goals in their IEPs (Individualized Education Plan).

Research Question Summation

To summarize, in terms of the research questions at the core of this dissertation, the overall objectives included (1) determining the relationship between student achievement and teacher effectiveness at the school level, (2) exploring this same relationship while controlling for teacher factors, and (3) examining this relationship while taking into account school factors. As such, the overarching goal was to determine the relationship between students standardized testing outcomes—achievement—and teacher APPR ratings—effectiveness—at the school level, particularly while controlling for student characteristics and other influential factors, such as free and reduced lunch or whether or not the student population was economically disadvantaged as a whole.

In both cases of ELA and math standardized testing results, APPR ratings were positively correlated to student achievement. After controlling for all other factors of consideration, the average percentage of teachers who were rated as *effective* or *highly effective* had no statistically significant association on the variable of student achievement (p=.905 and .864 for ELA and math, respectively). This leads to a few possibilities. There may be other variables not accounted for in this study that have a greater association to student achievement, such as teacher preparation programs, parental involvement, professional development, and curriculum alignment to the state standards. Also, the instruments used to measure teacher effectiveness and student achievement may not be the best indicators of teacher effectiveness and student achievement.

The second research question explored the relationship between teacher qualifications and student achievement. Teaching experience, the proportion of teachers with three years of experience or less, or the teachers' level of educational attainment proved not to be statistically

significant. In fact, when looking at ELA outcomes, the *p*-value was found to be .409 and .917 for the variables of experience and education, respectively, while presenting as $p = .958$ and $p = .636$ for these factors in terms of the math testing outcomes. Even more interesting, as the average number of teachers with three years or less experience and a master's or doctorate increased, the subsequent influence on student test outcomes for ELA was negative. This indicated that the average percentage of teachers with three years or less experience and greater education levels had a negative association to student achievement or no association, at all, if the significance level was considered.

Once again, this leads to the question of whether experience and education are reliable indicators of teacher effectiveness. Because the latter seems unlikely, a third possibility is that other factors may be influencing student achievement in the state of New York—a factor that supersedes even the influence of teachers themselves.

The final research question examined the extent of the relationship between teacher effectiveness and student achievement, while taking into account school factors. While all teacher-related factors seemed to have no association to student achievement, factors related to school profiles, or the composition of the student population, were the only variables found to have any relevant significance. More specifically, the average number of students qualifying for free lunch and reduced lunch, a lower income or low SES, and students with disabilities resulted in a statistically significant and negative association to student achievement.

Ultimately, the findings within this study represent similar findings to those of prior studies when it comes to the relationship between teacher effectiveness and student achievement when not controlling for other variables. This includes the 2017 study authored by Johnson in which it was found that an increase in teacher ratings produced a positive association with

student progress. Similarly, the earlier MET study found that teacher composite scores accurately predicted student performance, as indicated by state standardized test outcomes (Bill & Melinda Gates Foundation, 2013). Further, students assigned to *effective* teachers performed better than expected, when compared to students assigned to *less effective* teachers and performed below expectations as a function of it. Finally, Papay (2012) is a frequently cited study in which teacher evaluation ratings were also found to have a definitive relationship with student achievement. The findings of this study led to many questions related to the schools, the teachers, and the students in New York State, which informs research implications of this study.

Implications

When looking at the implications of these findings, some of the most significant may be applicable within the field of education itself. This includes the way in which linking educator effectiveness and student achievement play a role in how teachers are assigned and hired. This is partially a function of teacher ratings and standardized student tests as tools for measuring student achievement and teacher effectiveness.

Even when taking teachers' experience or educational levels into consideration, there was little empirical connection, if any, found between variables that were typically associated with teacher effectiveness and student achievement. This brings to light some interesting questions, considering that APPR ratings were shown to influence student achievement.

However, when none of the teacher-related variables analyzed were found to have any notable influence on student achievement, new revelations emerge. First and foremost, if none of these factors reflect on student achievement, what factors may be influential? This is a particularly insightful question for the field of education, when student achievement is the end goal of teaching and quality teachers are often selected according to their level of experience and

education, as well as retained according to their APPR ratings or sufficient performance on other annual reviews.

These emerging revelations also illuminate implications that are applicable to the field of educational research. Specifically, the field of educational research often involves a focus on the overarching role of teachers, the defining characteristics of quality teaching, and how it influences student outcomes. The findings of this study certainly warrant further attention by researchers, while also more closely examining the previously mentioned areas of inquiry. Additional implications relevant to the research field include a comprehensive assessment, or perhaps reassessment of how well standardized testing represents student achievement, as well as inquiry into the accuracy of teacher ratings, particularly APPR ratings, as a measure of teacher effectiveness. Finally, the relationship between teacher effectiveness and student achievement should be reassessed, focusing on the function of standardized test outcomes and teacher ratings as the variables used in operationalizing these concepts.

It is important to note here that teacher effectiveness scores (APPR) and student achievement scores (NYS ELA and math) are not necessarily destined to correlate. Multi-tiered systems of observations and assessments are designed to evaluate different things. It is the combination of the different layers that should be used to produce overall performance levels for teachers and students. The use of the outcomes from the various assessments can then be used to drive instructional programs for students and professional development programs for teachers.

In addition, the findings of this study also shed light on policy implications including the need to more thoroughly research educational policy and policy decisions related to teacher effectiveness. Historically, many of the goals inherent in the formulation of related policies has led to a greater reliance on standardized testing as a means of evaluating teacher effectiveness

and student learning (Beyer & Johnson, 2014). Yet in light of the findings presented within the context of this research, a greater understanding is required pertaining to how the underlying goals of policy are represented, formulated, and practically applied, while paying particular attention to the underlying mechanisms used to achieve these goals.

For example, the findings of this study suggested that teacher effectiveness (APPR scores) were a sufficient indicator of student achievement. However, when controlled with other variables, APPR ratings proved to be insignificant. This may not be a function of the ratings alone as the selected measure of teacher effectiveness. Instead, the flaw may be in how the ratings are utilized and applied for these purposes. In fact, prior findings presented within the research of Marshall (2013) suggested that the use of teacher evaluation ratings, or any form of value-added assessment, should be subject to the inclusion of data from a three-year period in order to achieve accurate results. This is but one area of policy research that should be studied further, not only examining how ratings are implemented in their practical application, but also how policy should incorporate these findings to ensure accuracy in results and achieve the intended outcomes that motivated the policy in the first place.

New York Specific Implications

Because the state of New York is the context for this dissertation, there are several implications for the educational system within the state or at least applicable to the districts included within this study. First and foremost, policymakers, the Department of Education, and other stakeholders should reevaluate how teachers are assessed and reconsider the accuracy of APPR ratings in identifying effective versus ineffective teachers. More specifically, the underlying goal of current legislation was to hold teachers accountable for student performance by formulating an evaluative system that linked teacher effectiveness to student achievement

(Ciaccio et al., 2017). However, the findings within this study reveal an evaluative system for teachers that bears no empirical connection to student achievement, thereby minimizing the current methods of assessment as (1) an accurate representation of faculty effectiveness, (2) an accurate reflection of the subsequent influence on student achievement, and (3) a valid means of promoting teacher accountability. This also reaffirms the assertions formerly set forth by Moldt (2016) which reported that educators found the law was not effective at improving accountability or instructional practices.

Last but not least, the findings of this study may have strong implications at the individual level, influencing teachers as well as the students they teach—particularly those students within the State of New York education system. In regards to teachers, their annual reviews may influence their ongoing employment (tenure), pay rate, or even institutional status. Educators may also fail to grow or improve in their teaching strategies, because of the inaccurate feedback produced from insufficient evaluative tools resulting in the opportunity to provide and or participate in professional development. The enthusiasm, attitude, and motivation level of teachers influences the attitude and motivational levels of students, thereby potentially promoting or even deterring student achievement and enthusiasm for learning.

Limitations

Several limitations of the study should be noted. Among these, although the models made an effort to account for many influencing factors in student achievement, accounting for all potential influential or extraneous variables, in all probability, may not be feasible. In addition, the methods of analyses involving correlational relationships may demonstrate associations between variables. There are inevitably other factors for consideration that were not accounted for within the confines of this study, such as teacher professional development, parental

involvement, curriculum, and teacher preparation programs.

Finally, there is the potential for confounding factors that are a product of the demographic population or geographic location. The sample size was also a limitation, representing schools, teachers, and students from only five of the 62 counties in New York. The sample size covers school districts from rural, urban, suburban, and city school districts in New York State, which encompasses a diverse student and teacher population. The school districts in the sample represented some of the wealthiest school districts in New York State to some of the neediest school districts. While the sample sought to cover a cross-section of the educational environment in New York State, there are still some demographics left to be examined.

Recommendations and Future Areas of Study

The suggestions for future areas of study also pose implications applicable to the education system in the state of New York. Future studies should be undertaken that reassess the utility of the instruments used for measuring the variables of interest in this study. This includes the use of standardized tests to measure student achievement, as well as the APPR ratings, for evaluating teacher effectiveness. This should be supplemented with studies that comparatively assess the accuracy of value-added assessments and the assertion that these evaluations should be implemented with at least three years of data for genuine accuracy (Marshall, 2013). If additional research endeavors reaffirm the findings realized within this study that APPR ratings are not an adequate indication of teacher effectiveness, then further research should be undertaken to identify more accurate tools of assessment. An effort should also be made to ensure that APPR scores are not the sole source for assigning, hiring, firing, and retaining teachers if the ultimate goal is student achievement.

In each of the aforementioned cases, the schools within this study, as well as the state of

New York educational system as a whole should implement efforts at finding answers to the inquiries mentioned, as well as facilitate additional studies that are focused on the New York State student population and the predominant factors that affect student achievement. This is a particular area of interest, considering the varying teacher-related factors that were tested within the context of this study and were found to have no significant influence on student outcomes when it is logical to assume that they would. As a result, further study is warranted to explore and identify what is occurring within the New York student population that is undermining students' ability to achieve overall and negating the influence of teacher-related factors as a whole. Special attention should be directed toward the effectiveness of faculty and the attention invested toward students with disabilities, as well as household characteristics and other factors that are associated with the achievement of students from lower income households. Once a possible causation is identified, this should be complemented by the formulation of strategies to mitigate the negative influence of the underlying causative mechanism, followed by the development of policy that will support the changes necessary.

References

Abendroth, M., & Porfilio, B. J. (2015). *Understanding neoliberal rule in K-12 schools*: *Educational fronts for local and global justice* (Vol. 1). Charlotte, NC: Information Age.

Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory Into Practice*, *42*(1), 18–29.

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis*, *39*(1), 54–76.

Alexander, E. (2016). *Teacher evaluation: The relationship between performance evaluation ratings and student achievement* (Doctoral dissertation, Liberty University). Retrieved fromhttps://pdfs.semanticscholar.org/9a66/3846b61e18d19403ffcf763f570ee2c63b6a.pdf

Baker, B., Barton, P., Darling-Hammond, L., Haertel, E., Ladd, H., Linn, R., Shepard, L. (2010). Economic Policy Institute Briefing Paper. *Problems with the use of student test scores to evaluate teachers.* Retrieved from http://www.epi.org/publication/bp278/

Baker, B., Oluwole, J., & Green, P. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the race-to-the-top era. *Education Policy Analysis Archive*, *21*(5), 1–68.

Beatty, A. S., Educational Testing Service, & National Center for Education Statistics. (1996). *NAEP 1994 U.S. history report card: Findings from the National Assessment of Educational Progress*. Washington, D.C: Office of Educational Research and Improvement, U.S. Dept. of Education.

Berends, M. (2004). In the Wake of *A Nation at Risk*: New American Schools' private sector school reform initiative. *Peabody Journal of Education*, *79*(1), 130–163.

Berliner, D. (2014). Exogenous variables and value-added assessments: A fatal flaw. *Teachers College Record*, *116*(1), 1–17.

Berliner, D. C. (2013). Problems with value-added evaluations of teachers? Let me count the ways! *Teacher Educator*, *48*(4), 235–243.

Beyer, B. M., & Johnson, E. S. (2014). *Special programs & services in schools: Creating options, meeting needs*. Lancaster, PA: Destech.

Bill and Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. Policy and Practice Brief, MET Project, Retrieved from httphttps://k12education.gatesfoundation.org/resource/ensuring-fair-and-reliable-measures-of-effective-teaching-culminating-findings-from-the-met-projects-three-year-study/

Bitler, M.P., Corcoran, S. P., Domina, T., Penner, E. K., & Society for Research on Educational Effectiveness (SREE). (2014). Teacher Effects on Student Achievement and Height: A Cautionary Tale. In *Society for Research on Educational Effectiveness*. Society for Research on Educational Effectiveness.

Bjork, C. (2015). *High-stakes schooling—what America can learn from Japan's experiences*. Chicago, IL: The University of Chicago Press.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 416–440.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2011). Teacher layoffs: An empirical illustration of seniority versus measures of effectiveness. *Education Finance and Policy*,*6*(3), 439–454.

Bozkurt, A., Akgun-Ozbek, E., Yilmazel, S., Erdogdu, E., Ucar, H., Guler, E., & Dincer, G. D. (2015). Trends in distance education research: A content analysis of journals 2009–2013. *The International Review of Research in Open and Distributed Learning*, *16*(1).

Camerino, O., Castañer, M., & Anguera, T. M. (Eds.). (2014). *Mixed methods research in the movement sciences: Case studies in sport, physical education, and dance* (Vol. 5). New York, NY: Routledge.

Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average* (2nd ed.). Daniels, WV: Friends of Education.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, *24*(4), 305-331.

Chetty, R., Friedman, J., Hilger, N., Saez, E., Schanzenbach, D., & Yagan, D. (2010). *How does your kindergarten classroom affect your earnings? Evidence from project star*. Bloomington, IN: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w16381.pdf

Chetty, R., Friedman, J., & Rockoff, J. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*. NBER Working Paper No. 17699.Retrieved from http://www.equality-of-opportunity.org/assets/documents/teachers_wp.pdf

Ciaccio, A., DeMaio, R., Flynn, A., Hanssler, S., Malone, M., Mare, S., Short, V. (2017). Tying teacher evaluation to student test performance in New York State. *Hofstra Law Student Works*, *11*. Retrieved from http://scholarlycommons.law.hofstra.edu/hofstra_law_student_works/11

Cohen, P., West, S. G., & Aiken, L. S. (2014). *Applied multiple regression/correlation analysis for the behavioral sciences*. New York, NY: Routledge.

Cramer, E., Little, M. E., & McHatton, P.A. (2018). Equity, equality, and standardization: Expanding the conversations. *Education and Urban Society*, *50* (5), 483–501.

Creswell, J. (2015). *Research design: Qualitative, quantitative and mixed methods approaches.* NY, New York: Sage.

Darling-Hammond, L. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement.* Williston, VT: Teachers College.

Darling-Hammond, L. (2014). One piece of the whole: Teacher evaluation as part of a comprehensive system for teaching and learning. *American Educator*, *38*(1), 4–13.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, *93*(6), 8–15.

Desilver, D. (2017). U.S. students' academic achievement still lags that of their peers in many other countries. *Pew Research Center*. Retrieved from https://www.pewresearch.org/fact-tank/2017/02/15/u-s-students-internationally-math-science/

Domanico, R., & Manhattan Institute for Policy Research. (2019). *Lift the Cap: Why New York City Needs More Charter Schools. Issue Brief. Manhattan Institute for Policy Research. Manhattan Institute for Policy Research*. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=eric&AN=ED594220&site=eds-live

Egalite, A. J., Kisida, B., & Winters, M. A. (2015). Representation in the classroom: The effect of own-race teachers on student achievement. *Economics of Education Review*, *45*, 44–52.

Forman, K. & Markson, C. (2015). Is "effective" the new "ineffective"? A crisis with the New York state teacher evaluation system. *Journal for Leadership and Instruction*. Retrieved from https://files.eric.ed.gov/fulltext/EJ1080699.pdf

Franco, M. S., & Seidel, K. (2014). Evidence for the need to more closely examine school effects in value-added modeling and related accountability policies. *Education and Urban Society*, *46*(1), 30–58.

Freedheim, D. K. (2003). *Handbook of psychology* (Vol. 1): *History of psychology*. New York, NY: John Wiley & Sons.

Friedman, K. (2018). What is the Stanford Achievement Test (SAT)? *LA Tutors*. Retrieved from http://www.latutors123.com/2018/01/05/what-is-the-stanford-achievement-test-sat10/

Fryer, R. G. (2013). Teacher incentives and student achievement: Evidence from New York City public schools. *Journal of Labor Economics*, *31*(2), 373–407.

Fuller, B., & Henne, M. K. (2008). *Strong states, weak schools: The benefits and dilemmas of centralized accountability*. Bingley, United Kingdom: Emerald.

Good, T. L. (2008). *21st century education: A reference handbook*. Los Angeles, CA: Sage.

Green, P., Baker, B., & Oluwole, J. (2012). The legal and policy implications of value-added teacher assessment policies. *Brigham Young University Education & Law Journal*, (1), 1–29.

Grissom, J. A., & Youngs, P. (2016). *Improving teacher evaluation systems: making the most of multiple measures.* Teachers College Press.

Haghighat, M., Abdel-Mottaleb, M., & Alhalabi, W. (2016). Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition. *IEEE Transactions on Information Forensics and Security*, *11*(9), 1984–1996.

Hamilton, L. S., & Koretz, D. M. (2002). Tests and their use in test-based accountability systems. In L. S. Hamilton., B. M. Stecher, & S. P Klein (Eds.), *Making sense of test-based accountability in education* (pp. 13–49). Santa Monica, CA: Rand.

Haney, W. (2002). Lake Woebe guaranteed: Misuse of test scores in Massachusetts, Part I. *Education Policy Analysis Archives*, *10*, 24–32.

Hanushek, E. A., & Raymond, M. E. (2004). *Does school accountability lead to improved student performance*? Working Paper No. 10591. Cambridge, MA: National Bureau of Economic Research.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*(2), 371–406.

Hollins, E. R. (2015). *Rethinking field experiences in preservice teacher preparation: Meeting new challenges for accountability*. New York, NY: Routledge.

Hursh, D. (2007). Assessing No Child Left Behind and the rise of neoliberal education policies. *American Educational Research Journal*, *44* (3), 493–518.

Ingersoll, R.M., & Collins, G. J. (2017). Accountability and Control in American Schools. *Journal of Curriculum Studies*, 49(1), 75–95.

Johnson, A. (2017). *The relationship between teacher practice and student performance* (Dissertations and Theses, Seton Hall University). Retrieved from https://scholarship.shu.edu/dissertations/2235

Kane, T. J., Staiger, D. O., Grissmer, D., & Ladd, H. F. (2002). Volatility in school test scores: Implications for test-based accountability systems. *Brookings Papers on Education Policy*, (5), 235–283.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2010*). Identifying effective classroom practices using student achievement data*. Cambridge, MA: National Bureau of Economic Research.

Keddie, A. (2015). Student voice and teacher accountability: Possibilities and problematics. *Pedagogy, Culture, and Society, 23*(2), 225–244.

Kimball, S. M. (2002). Analysis of feedback, enabling conditions and fairness perceptions of teachers in three school districts with new standards-based evaluation systems. *Journal of Personnel Evaluation in Education*, *16*(4), 241–268.

Klem, A. M., & Connell, J. P. (2004). Relationships matter: Linking teacher support to student engagement and achievement. *Journal of School Health*, *74*(7), 262–273.

Lauen, D. L., & Gaddis, M. (2012). Shining a light or fumbling in the dark? The effects of NCLB's subgroup-specific accountability pressure on student performance. *Educational Evaluation and Policy Analysis*, *34* (2), 185–208.

Lee, A. M. I. (2014). Common Core State Standards: What you need to know. *Understood.* Retrieved from https://www.understood.org/en/school-learning/partnering-with-childs-school/tests-standards/common-core-state-standards-what-you-need-to-know

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *29*(2), 4–16.

Linn, R. L. (2015). Test-based accountability: University of Colorado at Boulder Center for Research on Evaluation, Standards and Student Testing. *Educational Researcher*, *35*, 15–26.

Madaus, G. F. (2013). *The courts, validity, and minimum competency testing*. New York, NY: Springer Science and Business Media.

Marshall, K. (2013). *Re-thinking teacher supervision and evaluation: How to work smart, build collaboration, and close the achievement gap* (2nd Ed.). New York, NY: Jossey Bass.

Martin, D. J., & Loomis, K. S. (2013). *Building teachers: A constructive approach to introducing education*. Boston, MA: Cengage.

Marzano, R. J., Frontier, T., & Livingston, D. (2011). *Effective supervision: supporting the art and science of teaching.* ASCD.

Mathison, S., & Ross, E. W. (2013). The hegemony of accountability in schools and universities. *Workplace*, *9*, 88–102.

McCombs, J. S., Kirby, S. N., & Mariano, L. T. (2010). *Ending social promotion without leaving children behind: The case of New York City*. Santa Monica, CA: Rand.

Medlock, I. Y. W. (2017). *Teacher evaluation ratings and student achievement: What's the connection?* (Doctoral dissertation, Wingate University). ProQuest, LLC. Retrieved from http://www.proquest.com/en- US/products/dissertations/individuals.shtml

Mertler, C. A. (2007). *Interpreting standardized test scores: Strategies for data-driven instructional decision making*. London, England: Sage.

Miner, B. (2004). Seed money for conservatives. *Rethinking Schools*, *18*(4), 9–11.

Miner, B. (2005). Testing companies mine for gold. *Rethinking Schools*, *19*(2), 5–7.

Moldt, SR (2016 Making the grade: A ground-level analysis of New York State's teacher performance review under APPR. *Brigham Young University Education and Law Journal* (2): 217–262.

New York State Education Department (NYSED). (2019). *Guidance on New York's Annual Professional Performance Review Laws and Regulations.* Albany, NY.

New York State Education Department (NYSED) Archives (2015–2016). Retrieved from https://data.nysed.gov

Nichols, S. L. (2007). High-stakes testing: Does it increase achievement? *Journal of Applied School Psychology*, *23*(2), 47–64.

No Child Left Behind Act (NCLB) of 2001. (P.L.107–110 [20 U.S.C. 7801]). [Google Scholar]

Odden, A. (2014). Lessons learned about standards-based teacher evaluation systems. In *Assessing Teacher, Classroom, and School Effects* (pp. 130–141). New York, NY: Routledge.

Papay, J. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*(1), 123–141.

Perullo, D., & Princeton Review (Firm). (2003). *Roadmap to 4th grade English language arts*. New York, NY: Random House.

Ronfeldt, M., Farmer, S. O., McQueen, K., & Grissom, J. A. (2015). Teacher collaboration in instructional teams and student achievement. *American Educational Research Journal*, *52*(3), 475–514.

Robinson, S. B. (2020). Teacher Evaluation: Why it matters and how we can do better. Retrieved from https://www.frontlineeducation.com/teacher-evaluation/

Ruiz-de-Velasco, J., (2005). Performance-based school reforms and the federal role in helping schools that serve language-minority students. In A. Valenzuela (Ed.), Leaving children behind: How "Texas-style" accountability fails Latino youth (pp. 33–55). Albany: State University of New York Press.

Şahin, F., & Levent, F. (2015). Examining the methods and strategies which classroom teachers use in the education of gifted students. *The Online Journal of New Horizons in Education*, *5*(3), 73–82.

Salkind, N. J., & Rasmussen, K. (2007). *Encyclopedia of measurement and statistics* (Vol. 1). Thousand Oaks, CA: Sage.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2010). *Assessment in special and inclusive education*. Belmont, CA: Cengage.

Sanders, W. L., Wright, S. P., & Horn, S. P. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, *11*(1), 57–67.

Schneller, P. (2017). Capitalism and public education in the United States. *Education Policy, Reforms and School Leadership*, (15), 101–106

Seifert, E. H., & Vornberg, J. A. (2002). *The new school leader for the 21st century, the principal*. Lanham, MD: Scarecrow Press.

Sloat, E., Amrein-Beardsley, A., Tenpe, A. Z., & Sabo, K. E. (2018). The TAP system for teacher and student advancement: A (questionable) system of teacher accountability and professional support. *Editorial Review Board*, *2018*, 3.

Steele, J. L., Hamilton, L. S., & Stecher, B. M. (2011). *Using student performance to evaluate teachers*. Santa Monica, CA: Rand.

Sunderman, G. L., Kim, J. S., & Orfield, G. (2005). *NCLB meets school realities: Lessons from the field*. Thousand Oaks, CA: Corwin Press.

Taylor, E., & Tyler, J. (2012). Can teacher evaluation improve teaching? *Education Next*, *12*(4), 78–84.

Traub, R. E., & Canadian Education Association. (1994). *Standardized testing in Canada: A survey of standardized achievement testing by ministries of education and school boards*. Toronto, Ontario: Canadian Education Association.

Troia, G. A., & Olinghouse, N. G. (2013). The Common Core State Standards and evidence-based educational practices: The case of writing. *Grantee Submission*, *42*(3), 343–357.

Tsoi-A, R., & Bryant, F. (2015). *College preparation for African American students: Gaps in the high school educational experience*. Washington, DC: Clasp.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, *73*(1), 89–122.

White, G. W., Stepney, C. T., Hatchimonji, D. R., Moceri, D. C., Linsky, A. V., Reyes-Portillo, J. A., & Elias, M. J. (2016). The increasing impact of socioeconomics and race on standardized academic test scores across elementary, middle, and high school. *American Journal of Orthopsychiatry*, *86*(1), 10–23.

Wiggins, G. (2011). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan*, *92*(7), 81–93.

Yüksel, P., & Yıldırım, S. (2015). Theoretical frameworks, methods, and procedures for conducting phenomenological studies in educational settings. *Turkish Online Journal of Qualitative Inquiry*, *6*(1), 1–20.